High-dimensional Modeling via Nonconcave Penalized Likelihood and Local Likelihood (Dissertation)

Runze Li
Department of Statistics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3260, USA

Advisors: Drs. Jianqing Fan & James S. Marron

April 19, 2000

High-dimensional Modeling via Nonconcave Penalized Likelihood and Local Likelihood

by
Runze Li

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics.

Chapel Hill

2000

Approved by

Advisor: Dr. Jianqing Fan

Advisor: Dr. James S. Marron

Reader: Dr. Gordon Simons

Reader: Dr. Indra M. Chakravarti

Reader: Dr. Richard L. Smith

ABSTRACT

Runze Li: High-dimensional Modeling via Nonconcave Penalized Likelihood and Local Likelihood

(Under the direction of Drs. J. Fan and J. S. Marron)

This dissertation consists of two parts: model selection via nonconcave penalized likelihood and statistical modeling via local likelihood. In Part I of this dissertation, a few variable selection procedures via penalized likelihood are proposed. The proposed methods select variables and estimate coefficients simultaneously. Hence it is easy to construct confidence intervals for estimated parameters. A new algorithm is proposed for optimizing high-dimensional nonconcave functions. The proposed ideas are widely applicable. They are readily applied to a variety of parametric models and semi-parametric models. They can also be applied to nonparametric modeling by using wavelets and splines. It has been demonstrated how the rates of convergence for the proposed estimators depend on the regularization parameter. Oracle properties have been established for the proposed variable selection procedures in this dissertation. In Part II of this dissertation, an efficient estimation procedure is proposed to estimate coefficient functions in varying-coefficient models. The asymptotic normality of the resulting estimators is established. The standard error formulae for estimated coefficients are derived and empirically tested. A goodness-of-fit test technique, based on a nonparametric maximum likelihood ratio type of test, is also proposed to detect whether certain coefficient functions in a generalized varying-coefficient model are constant or whether any covariates are statistically significant in the model. The null distribution of the test is estimated by a conditional bootstrap method. To reduce computational burden, a one-step Newton-Raphson estimator is proposed and implemented. It is shown that the resulting one-step procedure can save computational cost by a factor of tens without deteriorating its performance, both asymptotically and empirically. The SiZer map has been developed by Chaudhuri and Marron (1999) as a methodology for finding which features in noisy data are strong enough to be distinguished from background noise. An extension of the SiZer map with more efficiency in distinguishing features is proposed in Chapter 7.

ACKNOWLEDGEMENTS

I want to express my appreciation to Professor Gordon Simons, chairman of my dissertation committee, and Professors Indra M. Chakravarti and Richard L. Smith, who have served on my dissertation committee.

I am very grateful to my advisors, Professors Jianqing Fan and James S. Marron, for their inspiring guidance and enthusiastic support in this research. My indebtedness to them is more than words can tell. Their superb intuition, their broad knowledge, their availability for in-depth discussion, and their constant encouragement throughout this work have all been indispensable for whatever I have achieved. The years that we worked together while in Chapel Hill are treasured memories of mine.

Finally, I would like to thank Dr. Fan's family and Dr. Marron's family for their hospitality, to thank my parents, my parents-in-law and my wife for their love, support and understanding.

TABLE OF CONTENTS

1 Introduction and Background for Part I **1**

1.1 Variable Selection for Linear Regression Models 2

1.2 Variable Selection in Nonparametric Regression 3

1.2.1 Spline method 4

1.2.2 Orthogonal series based methods and thresholding rules 5

1.3 Organization of Part I 7

2 Variable Selection Via Nonconcave Penalized Likelihood **11**

2.1 Penalized Least-squares and Variable Selection 12

2.1.1 Thresholding and variable selection 12

2.1.2 Penalized least-squares and variable selection 13

2.1.3 Smoothly clipped absolute deviation penalty 16

2.1.4 Performance of thresholding rules 18

2.2 Variable Selection via Penalized Likelihood 20

2.2.1 Penalized least-squares and likelihood 20

2.2.2 A unified algorithm 21

2.2.3 Standard error formula 23

2.2.4 Testing convergence of the algorithm 24

2.3 Examples 24

2.4 Simulations 28
2.4.1 Prediction and model error 28
2.4.2 Selection of thresholding parameters 29
2.4.3 Simulation study 30
2.5 Proofs 34
2.6 Conclusion 36

3 Theoretical Results and Oracle Properties 38

3.1 Introduction 38
3.2 Penalized Least-squares Estimators 40
3.3 Robust Penalized Least-squares Estimators 44
3.4 Penalized Likelihood Estimators 48
3.5 Discussion and Conclusion 56

4 Variable Selection in Survival Data Analysis 58

4.1 Introduction 58
4.2 Proportional Hazards Models 60
4.2.1 Cox's proportional hazard model and penalized partial likelihood 61
4.2.2 Sampling properties and oracle properties 62
4.3 Frailty Model 64
4.4 Simulation Studies and Applications 66
4.4.1 Local quadratic approximations and standard errors 66
4.4.2 Selection of thresholding parameters 67
4.4.3 Prediction and model error 67
4.4.4 Simulations 68
4.4.5 Application 71
4.5 Proofs 75

4.6 Conclusion 78

5 Introduction and Literature Review for Part II **79**

5.1 Conventional Kernel Regression 79

5.2 Local Polynomial Regression 81

5.3 Local Likelihood Approach 82

5.4 Organization of Part II 85

6 Generalized Varying Coefficient Models **87**

6.1 Introduction 87

6.2 Estimation of Varying Coefficients 89

6.2.1 Local MLE 89

6.2.2 One-step local MLE 90

6.2.3 Standard errors 90

6.2.4 Implementation of one-step local MLE 91

6.3 Main Results 92

6.4 Hypothesis Testing for Varying Coefficients 98

6.5 Simulations and Applications 99

6.5.1 Logistic Regression 99

6.5.2 Poisson regression 103

6.5.3 Real-data examples 107

7 SiZer Map Based on Local Likelihood **113**

7.1 Introduction 113

7.2 Generalized Linear Models and Quasi-likelihood Functions 116

7.3 Local Quasi-likelihood 117

7.4 SiZer Map Based on Local Quasi-likelihood 118

7.4.1 One-step local quasi-likelihood estimator 118
7.4.2 Numerical implementation of binned methods 120
7.4.3 SiZer map 121
7.5 Simulation and Application 123
7.5.1 Poisson regression 123
7.5.2 Logistic regression 124
7.5.3 Application 127
7.6 Discussion 130

Bibliography **131**

LIST OF TABLES

2.1 Estimated coefficients and standard errors for Example 2.3.1 26

2.2 Estimated coefficients and standard errors for Example 2.3.2 28

2.3 Simulation results for linear regression models 31

2.4 Standard deviations of estimators for linear regression models ($n = 60$) 31

2.5 Simulation results for robust linear models 33

2.6 Standard deviations of estimators for robust regression models 33

2.7 Simulation results for logistic regression 33

2.8 Standard deviations of estimators for logistic regression 34

2.9 Simulation results for Poisson log-linear regression 34

2.10 Standard deviations of estimators for Poisson log-linear models 35

4.1 Simulation results for Cox's proportional hazards model 70

4.2 Standard deviations in Cox's proportional hazards models ($n = 100$) 70

4.3 Comparisons between the proposed approach and pseudo likelihood approach . 71

4.4 Simulation results for frailty model 72

4.5 Standard deviations in frailty model ($G = 100,\ n = 2$) 72

4.6 Results for nursing home data (Cox's model) 73

4.7 Results for nursing home data (frailty model) 74

5.1 Pointwise asymptotic biases and variances 81

6.1 Bivariate summary of simulation results for logistic regression model 101
6.2 Standard deviations of estimators for logistic regression model 101
6.3 Six empirical percentiles for logistic model 102
6.4 Bivariate summary of simulation output for Poisson regression model 103
6.5 Standard deviations of estimators for Poisson regression model 105
6.6 Six empirical percentiles for Poisson model 105

LIST OF FIGURES

2.1 Plot of thresholding functions 14

2.2 A plot of $\theta + p'_\lambda(\theta)$ 15

2.3 Risk functions of proposed procedures under quadratic loss. 17

2.4 Four penalties $p_\lambda(\theta)$ and their quadratic approximations 22

2.5 Boxplots of relative model errors 36

6.1 Simulation results for Example 6.5.1 with sample size 400 104

6.2 Simulation results for Example 6.5.2 with sample size 200 106

6.3 Scatterplot of environmental data and cross-validation functions 109

6.4 Estimated coefficient functions for environmental data set 110

6.5 Pearson's residuals and autocorrelation coefficients 110

6.6 The estimated density of T 111

6.7 Estimated coefficient functions for the burn data set 112

7.1 True signal in Example 7.1.1 115

7.2 SiZer maps for sample size $n = 500$ 115

7.3 SiZer maps for sample size $n = 200$ 116

7.4 SiZer maps for Poisson regresion with $n = 500$ 125

7.5 SiZer maps for Poisson regression with sample size $n = 200$ 126

7.6 SiZer maps for logistic regression with sample size $n = 200$ 128

7.7 SiZer maps for environment data 129

7.8 SiZer maps for conditional probability 129

Chapter 1

Introduction and Background for Part I

Regression is one of the most useful techniques in statistics. Consider the $(d+1)$-dimensional data $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$, which form an independent and identically distributed sample from a population $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d-dimensional random vector and Y is a random variable. Of interest is to estimate the regression function $m(\mathbf{X}) = E(Y|\mathbf{X} = \mathbf{x})$. In other words, the data are regarded as realizations from the model:

$$Y = m(\mathbf{X}) + \varepsilon,$$

where ε is a random error with zero mean. In the context of linear regression, $m(\mathbf{x})$ is approximated by a linear combination of polynomials or other functions of $\mathbf{x}$. To attenuate possible excessive modeling biases, a large number of predictors are usually introduced at the initial stage of modeling, resulting in a high-dimensional linear regression model. In order to enhance model predictability and avoid collinearity problems, variable selection is fundamental to high-dimensional statistical modeling. Furthermore, regression equations with fewer variables have the appeal of simplicity, as well as an economic advantage in terms of obtaining the necessary information to use the equations. In addition, there is a theoretical advantage of eliminating irrelevant variables and in some cases, even variables that contain some predictive information about response variable. Variable selection is important not only for linear regression models, but also for other parametric models, such as robust linear regression model and generalized linear models, semiparametric models, e.g., Cox's proportional hazards model, and nonparametric regression models. A good variable selection procedure should result in an estimator

with three properties: (a) **unbiasedness**: the resulting estimator is unbiased when the true unknown parameter is large in order to avoid unnecessary modeling bias; (b) **sparsity**: the resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to be zero in order to reduce model complexity; (c) **continuity**: the resulting estimator is continuous in some sense in order to avoid instability in model prediction, see Chapter 2 for details. The traditional variable selection procedures, such as stepwise deletion and best subset selection, yield an unbiased and sparse estimator, but not a continuous estimator. On the other hand, ridge regression, a useful technique for improving least-squares estimators, results in a continuous estimator, but is not unbiased and not sparse. In Part I of this dissertation, a few new procedures of variable selection will be proposed. The proposed approaches have these three properties and can automatically select significant variables and estimate coefficient parameters simultaneously. They can improve stepwise deletion and best subset variable selection in two aspects: stability in model prediction and computational cost of best subset selection; and improve ridge regression in terms of reduction of model complexity and modeling bias.

1.1 Variable Selection for Linear Regression Models

Consider the usual linear regression model

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

where $\mathbf{y}$ is an $n \times 1$ vector and $\mathbf{X}$ is an $n \times d$ matrix. As in the traditional linear regression model, it is assumed that Y_i's are conditionally independent given the design matrix. The ordinary least-squares estimate is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The model may contain a number of irrelevant covariates in practice. To enhance predictability and to select significant variables, statisticians usually apply subset selection techniques, to improve the least-squares estimate.

Variable selection is an important topics in the context of linear regression. There is a large amount of literature on this subject. The monograph of Miller (1990) gives an excellent overview of this field. This section gives a brief summary of recent development in the field.

Many of the variable selection procedures in use are stepwise selection procedures. These procedures usually consist of two stages: Firstly choose a criterion of variable selection such as AIC (Akaike 1970, 1974), BIC (Schwarz 1978), Covariance Inflation Criterion (referred to

as CIC, Tibshirani and Knight, 1999), Mallows' C_p (Mallows, 1973), PRESS and adjusted R^2 among others; and then select an algorithm for finding subsets which fit well among algorithms: forward selection, stepwise regression, backward elimination, sequential replacement algorithms and the algorithm of exhaustive search over all possible subsets (referred to as best subset) among others. These procedures can be expensive in computation. Some techniques, such as grouping variables and leap and bound techniques, can be used for reducing the computational burden.

These procedures are practically useful. However, they ignore stochastic errors inherited in the previous stages of variable selection. Hence, their theoretical properties are somewhat hard to understand. Furthermore, it is known that best subset variable selection suffers from several drawbacks. The most severe one is its lack of stability as analyzed for instance by Breiman (1996).

In the situation in which there are a larger number of predictors, one prefers variable selectors which can select significant variables and estimate parameters simultaneously. Thresholding rules, such as hard and soft thresholding rules, have this property. There are many connections between thresholding rules and variable selection for linear regression model. In fact, when the design matrix is column-orthonormal, stepwise deletion and best subset variable selection are equivalent to the hard-thresholding rule with a proper thresholding parameter. The soft-thresholding rule forms the core of LASSO, proposed by Tibshirani (1996). The non-negative garrote, proposed by Breiman (1995), is a thresholding rule when the design matrix is orthogonal, and has been used to estimate the coefficients of wavelet transforms (Gao 1998). Best subset variable selection, LASSO and the non-negative garrote do not have the properties of unbiasedness, sparsity and continuity simultaneously. Their drawbacks can be repaired by the approaches proposed in Chapter 2.

1.2 Variable Selection in Nonparametric Regression

Variable selection also plays an important role in nonparametric regression. The monograph of Fan and Gijbels (1996) gives detailed background and excellent overview on various nonparametric regression techniques. A brief description of the spline method and the orthogonal series based method is given in this section in order to provide useful background for understanding significance of the proposed approaches in this dissertation and its various possible extensions.

Nonparametric regression techniques may be classified into two categories. One is to parameterize the regression function $m(x)$ locally, and the other one is to approximate the regression function globally. Two common methods of global approximation are the *spline approach* and the *orthogonal series based method*. See, for example, de Boor (1978), Wahba (1990) and Daubechies (1992). The spline approach allows possible discontinuities of derivative curves. The locations of discontinuity points, called knots, can be selected by the data via a *smoothing spline* method (see for example Wahba (1975)), a stepwise deletion method (see the work of Stone and his collaborators, for example Kooperberg and Stone (1991)) or Bayesian method (see for example Smith and Kohn (1996)). Another method of global approximation is to expand the regression function into orthogonal series, then choose a useful subset of the basis functions, and use them to approximate the regression function. This approach is called the *orthogonal series based method*. Local linear regression is to apply the linear regression model locally instead of increasing the number of parameters. That is for any given point x, model $m(\cdot)$ linearly around x and apply the linear regression technique to a fraction of data around x. This approach is termed the *local (linear) modeling approach*, which will be introduced in detail in Chapter 4 as background of Part II of this dissertation. See, for example, Fan and Gijbels (1996) for systematic study of the topic of local modeling and its applications. A very important issue for the global approximation methods is how to select a useful subset of the basis functions.

1.2.1 Spline method

Compared with polynomial regression, spline regression may yield a more flexible model. There is a large literature on the topic of smoothing splines. Schoenberg (1964), Reinsch (1967) and Kimeldorf and Wahba (1970, 1971) are early references on smoothing splines. de Boor (1978) gave practical implementations of spline techniques. Utreras (1980) and Li (1985, 1986) discussed the choice of smoothing parameters. Rice and Rosenblatt (1983) established asymptotic theory for smoothing splines. The theory of spline approximations and their statistical applications can be found in reference books, such as Eubank (1988) and Wahba (1990). Green and Silverman (1994) gives a comprehensive account on applications of spline techniques to various fields of statistics.

Consider the regression model

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1.2.1}$$

where ε_i's are independent and identically distributed with zero mean and common variance σ^2. The homoscedasticity of the model (1.2.1) is important to the development of spline smoothing techniques, although those techniques can be extended to a heteroscedastic model. In the full model, the regression function $m(\mathbf{x})$ may be too complicated to estimate. Two alternative models are the additive model $m(\mathbf{x}) = f_1(X_1) + \cdots + f_d(X_d)$ and the varying-coefficient model $m(\mathbf{x}) = a_1(X_1) + a_2(X_1)X_2 + \cdots + a_d(X_1)X_d$ defined in Chapter 6, which are simpler and useful nonparametric models for exploring fine features in the data.

For given knots, a set of spline basis functions can be constructed. Regression spline basis and B-spline basis (see de Boor (1978) for definition) are two popular spline bases. Usually, the B-spline basis is numerically more stable because the multiple correlation among the basis functions is small, but the regression spline basis has the advantage that deleting the basis function is the same as deleting the knots.

Let $\{B_1, \cdots, B_J\}$ be a set of spline basis functions, and approximate the regression function $m(\mathbf{x})$ by a linear combination of the basis functions, $\sum \theta_j B_j(\mathbf{x})$ say, which yields a high dimensional linear regression model. The least-squares approach may be applied for estimating the coefficients θ_j. After approximating the regression function, it remains very critical to select significant variables (terms in the expansion) to efficiently represent unknown functions. In a series of work by Stone and his collaborators (See Stone *et al* 1997), they modify traditional variable selection approaches to select useful spline subbases.

An alternative approach to automatic selection of knots is the smoothing spline. The idea of the smoothing spline is to find the estimation of $m(\cdot)$ by a penalized least squares regression instead of by a naive least squares regression. A penalty on roughness of $\hat{m}(\cdot)$ in the penalized least squares regression is used to avoid over-parametrization, and a smoothing parameter is introduced to control model complexity. The smoothing parameter can be chosen objectively by the data. One approach to select the smoothing parameter is via the minimization of the cross-validation. An improved version of this is the *generalized cross-validation*, proposed by Wabha (1977).

1.2.2 Orthogonal series based methods and thresholding rules

Orthogonal series based methods, including Fourier regression and wavelet-based methods, are alternative ways to approximate the regression function globally with a large number of unknown parameters. The orthogonal series method was introduced by Čencov (1962).

Rutkowski (1982) used this method to estimate a regression function nonparametrically. Wavelet transforms are a device for representing functions in a way that is local in both time and frequency domains. They have recently received a great deal of attention in applied mathematics, image analysis, signal compression, and many other fields of engineering. Good introductory references to this subject include Meyer (1990), Chui (1992), Daubechies (1992) and Strang (1989, 1993).

Wavelet-based methods, demonstrated by Donoho (1995) and Donoho and Johnstone (1994a, 1995, 1998), have many exciting statistical properties. An excellent overview on wavelet statistical estimation can be found in Donoho, Johnstone, Kerkyacharian and Picard (1995). The relationship between wavelet thresholding and local kernel smoothing methods is illuminatingly described by Hall and Patil (1995a, b) and an interesting comparison in terms of efficiency is given in Fan, Hall, Martin and Patil (1996, 1999). Useful reference books on statistical application of wavelets include Ogden (1997), Härdle, Kerkyacharian, Pickard and Tsybakov (1998) and Vidakovic (1999).

For simplicity of presentation, consider the one-dimensional canonical regression model:

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $x_i = \frac{i}{n}$ and ε_i is a sequence of independent and identically distributed Gaussian noise $\varepsilon \sim N(0, \sigma^2)$. The assumptions on the equispaced design and the Gaussian white noise are important for the development of the methodology, although the basic idea should in principal be applicable to non-canonical regression models. See for example Antoniadis and Fan (1999). For given n, (usually $n = 2^k$ for some integer k so that fast computation algorithms can be implemented), a set of orthonormal basis functions of R^n with desired properties can be constructed and $\mathbf{m} = (m(x_1), \cdots, m(x_n))^T$ can be represented as $\mathbf{W}\boldsymbol{\theta}$, where $\mathbf{W}$ is an $n \times n$ orthogonal matrix consisting of the orthonormal basis functions. Thus the canonical regression model is reduced to

$$\mathbf{y} = \mathbf{W}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

which is a high-dimensional linear regression model. The least-squares estimate for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}} = \mathbf{W}^T\mathbf{y}$.

In the case of Fourier analysis, one would expect that the energy of a small underlying signal $m(\cdot)$ at high frequencies is small (i.e. $|\theta_j|$ is small when j is large), that is, the information is compressed into low frequencies. Therefore use of the first N basis functions would be reasonable for modeling $m(\cdot)$, and N is regarded as a smoothing parameter. However, if such

primary information is not granted, then an ideal choice of the subset of basis functions consists of those having large coefficients $|\theta_j|$. To avoid expensive computation and accumulation of stochastic errors in subset selection, Donoho and Johnstone (1994a) suggested the use of hard- and soft-thresholding rules introduced by Bickel (1983):

$$\hat{\theta}_{H,j} = \hat{\theta}_j I\{|\hat{\theta}_j| > \delta\}$$

and

$$\hat{\theta}_{S,j} = \mathrm{sgn}(\hat{\theta}_j)(|\hat{\theta}_j| - \delta)_+,$$

where $\delta > 0$ is a thresholding parameter, $I\{|\hat{\theta}| > \delta\}$ is an indicator function, giving values 1 where $|\hat{\theta}| > \delta$ and 0 otherwise, and $\mathrm{sgn}(\cdot)$ denotes the sign function.

Donoho and Johnstone (1994) proposed using $\delta = \sqrt{2\log n}\sigma$, and referred to such a thresholding as *visual shrinkage* since it often produces the best visual pictures. The choice of this particular thresholding parameter is based on extreme value theory for the Gaussian white noise:

$$\{\sqrt{2\log(n)}\sigma\}^{-1} \max_{1\le i\le n} \varepsilon_i \to 1, \quad \text{a.s.}$$

The choice has the property of asymptotic optimality as described in Donoho and Johnstone (1994).

The *universal thresholding* suggested by Donoho and Johnstone (1994) is $\delta = \sqrt{2\log(n)}\hat{\sigma}$, where $\hat{\sigma}$ is the median absolute deviation of the wavelet coefficients at the highest resolution, divided by 0.6745. This is based on robust estimation of the standard deviation under the normal model. However, in practical implementation, one may also want to try other values of the thresholding parameters.

1.3 Organization of Part I

The first part of this dissertation consists of Chapters 1, 2, 3 and 4. Chapter 2 is devoted to developing a new approach to selecting significant variables for linear regression models. To attenuate possible excessive modeling biases, a large number of predictors are usually introduced at the initial stage of modeling. To enhance predictability and to select significant variables, statisticians usually apply three standard techniques, stepwise deletion, subset selection and ridge regression, to improve the least-squares estimate. The former two methods are variable selection techniques and the latter one is not. While they are practically useful,

these selection procedures ignore stochastic errors inherited in the previous stages of variable selections. Hence, their theoretical properties are somewhat hard to understand. In an attempt to automatically and simultaneously select variables, we propose a unified approach via penalized least-squares, retaining good features of both subset selection and ridge regression. The penalty functions have to be singular at the origin in order to produce sparse solutions (many estimated coefficients are zero), to satisfy certain conditions to produce continuous models (for stability of model selection) and to be bounded by a constant to produce nearly unbiased estimates for large coefficients. The bridge regression proposed in Frank and Friedman (1993) and the LASSO proposed by Tibshirani (1996, 1997) are members of the penalized least-squares family, though their associated L_p-penalty functions do not satisfy all of the three required properties above.

The penalized least-squares idea can naturally be extended to robust linear regression models and likelihood based models in various statistical contexts. The proposed approaches are distinguished from traditional methods (usually quadratic penalty) in that the penalty functions are symmetric, convex on $(0, \infty)$ (rather than concave as for the quadratic penalty in the penalized likelihood situation) and possess singularities at the origin. A few new penalty functions are introduced. These allow statisticians to select a penalty function to enhance the predictive power of a model and engineers to choose a penalty function to sharpen noisy images. Optimizing a penalized likelihood is very challenging, since the target function is a high-dimensional nonconcave function with irregularities. A new and generic algorithm is proposed for such an optimization problem. This yields a unified variable selection procedure. A standard error formula for estimated coefficients is obtained by using a sandwich formula, via the proposed iterative algorithm. Tests show that the formula is accurate enough for even when the sample size is very moderate. The proposed procedures are compared with various other variable selection approaches. The results indicate favorable performance of the newly proposed procedures.

The proposed penalized likelihood method can be applied readily to high-dimensional nonparametric modeling. After approximating regression functions using splines or wavelets, it remains very critical to select significant variables (terms in the expansion) to efficiently represent unknown functions. In a series of work by Stone and his collaborators (see Stone *et al* 1997), they modify traditional variable selection approaches to select useful spline subbases. It remains very challenging to understand the sampling properties of these data-driven variable selection techniques. Penalized likelihood approaches, outlined in Wahba (1990), Silverman

and Green (1994) and references therein, are based on a quadratic penalty. They reduce the variability of estimators via ridge regression. In wavelet approximations, Donoho and Johnstone (1994a) select significant subbases (terms in the wavelet expansion) via thresholding. The proposed penalized likelihood approach can be directly applied to these problems (see Fan and Antoniadis, 1999). Because we select variables and estimate parameters simultaneously, the sampling properties of such a data-driven variable selection method can be established.

Chapter 3 demonstrates how the rates of convergence for the penalized least-squares estimators, robust penalized least-squares estimators and the penalized likelihood estimators depend on the regularization parameter. It will be shown further the penalized likelihood estimators perform as well as the oracle procedure in terms of selecting the correct model, when the regularization parameter is appropriately chosen. In other words, when the true parameters have some zero components, they are estimated as 0 with probability tending to one, and the nonzero components are estimated as well as if the correct submodel were known. This improves the accuracy for estimating not only the null components, but also non-null components. In short, the penalized likelihood estimators work as well as if the correct submodel were known in advance. The significance of this is that the proposed procedures outperform the maximum likelihood estimator and perform well . This is very analogous to the super-efficiency phenomenon, for instance, Hodges example in page 405 of Lehmann (1983).

Chapter 4 will focus on variable selection for semi-parametric models in the context of survival data analysis. It will be shown that penalized likelihood for the Cox regression model is equivalent to penalized partial likelihood. The rates of convergence and oracle properties will be established for the proposed penalized partial likelihood estimators. Furthermore variable selection for Cox's proportional hazard frailty models will be studied. There are a number of papers concerned with the estimation problems of regression parameters in the recent literature. Lee *et al* (1992) suggested the use of pseudo-likelihood approach to estimate the regression coefficients when the number of members in each group is small. Nielsen *et al* (1992) applied the EM algorithm to the gamma-frailty model (see Section 4.3 below for definition). Sinha (1998) proposed a posterior likelihood method to the gamma frailty model. It seems that there is no satisfactory method for estimating regression coefficients in semi-parametric frailty models. Therefore a new approach to estimating the regression coefficients in Cox's proportional hazard frailty models is proposed, motivated by the idea of backfitting in the context of nonparametric regression. The proposed approach can be used to derive a unified variable selection method via nonconcave penalized likelihood for semi-parametric frailty model. Numerical simulations

are conducted to compare the performance of the proposed approach and the pseudo-likelihood estimator. The performance of the proposed approaches is also compared with that of best subset variable selection.

Chapter 2

Variable Selection Via Nonconcave Penalized Likelihood

As discussed in the last chapter, variable selection plays important roles in both parametric regression and nonparametric regression. This chapter focuses on the topic of variable selection in the context of parametric regression, including linear regression, robust linear regression and the generalized linear model.

There are strong connections between penalized least-squares methods and variable selection in linear regression models. When design matrices are orthonormal, the stepwise backward deletion and the best subset selection methods are equivalent to a hard-thresholding rule. The latter keeps an estimated coefficient intact when it exceeds a certain thresholding level, and sets it to zero otherwise. The hard thresholding estimator can be regarded as a solution to a penalized least-squares problem, as to be shown in Section 2.1. The hard thresholding estimator is discontinuous and this can be repaired by using the soft-thresholding rule, which sets small coefficients to 0 and shrinks the estimate by a constant. The latter is also a solution to the penalized least-squares with the L_1-penalty, on which the LASSO is based. Figures 2.1 (a) and (b) show that hard-thresholding and soft-thresholding functions. The hard-thresholding rule coincides with best subset variable selection when design matrices are orthonormal. In Breiman's (1996) terminology, best subset variable selection may be seen as a regularization technique. However, it is known that regularization by best subset variable selection suffers from several drawbacks, and the most severe one is its lack of stability as analyzed for instance by Breiman (1996). The hard-thresholding rule results in an unstable model in the sense that a

small change of data can lead to a very different model. This can create excessive variabilities in prediction. On the other hand, while the soft-thresholding rule is continuous, it always shifts an estimate by a constant. This would cause a lot of bias if the thresholding parameter is large. In the same spirit of Bruce and Gao (1997), in the discussion of a paper by Antoniadis (1999), Fan outlined a few thresholding rules which aim at improving the properties of both the hard and soft thresholding rules. These new rules can also be regarded as penalized least-squares. In particular, a smoothly clipped absolute deviation (SCAD) penalty function is proposed to improve the properties of the L_1 and the hard-thresholding penalty functions. This will be demonstrated in this chapter.

2.1 Penalized Least-squares and Variable Selection

Consider the general linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector and $\mathbf{X}$ is an $n \times d$ matrix. As in the traditional linear regression model, assume that y_i's are conditionally independent given the design matrix.

There are strong connections between thresholding rules and subset selection in linear regression models. To give more insight about various variable selection procedures, assume that the columns of $\mathbf{X}$ in (2.1.1) are orthonormal. Then the least-squares estimate in the full model is $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$, a part of the orthogonal transform of the vector $\mathbf{y}$. In this section, comparisons among various variable selection procedures will be conducted within the framework of penalized least-squares.

2.1.1 Thresholding and variable selection

The backward stepwise deletion algorithm in the linear models is to delete a variable, one at a time, with the smallest absolute t-value. For the orthonormal design matrix, this corresponds to deleting the variable with the smallest absolute value of estimated coefficient. When a variable is deleted, the remaining columns of the design matrix $\mathbf{X}$ are still orthonormal and the estimated coefficients remain unchanged. So in the second step, the algorithm deletes the variable that has the second smallest estimated coefficient in the full model. If the stepwise backward deletion is carried out m times, the remaining variables are those with the largest

$n - m$ values of $|\hat{\boldsymbol{\beta}}|$. This is equivalent to using a hard thresholding rule with a thresholding parameter between the m^{th} and $(m+1)^{th}$ order statistics of $|\hat{\boldsymbol{\beta}}|$.

The soft-thresholding rule can be viewed similarly. Denote $\mathbf{z} = \mathbf{X}^T\mathbf{y}$ and assume that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ in model (2.1.1). Then $\mathbf{z}$ is a multivariate normal random vector with independent components. This allows us to consider a Gaussian white noise model:

$$z_i = \theta_i + \varepsilon_i \quad \text{with } \varepsilon_i \sim N(0, \sigma^2) \quad \text{for } \ i = 1, \cdots, d. \tag{2.1.2}$$

Suppose that the θ's in (2.1.2) are sparse so that it is reasonable to assume that the prior for θ_i is a double exponential distribution with a scale parameter λ_1, where λ_1 is a hyperparameter. Then the Bayesian estimate of θ_i is the minimizer of

$$\frac{1}{2}\sum_{i=1}^{d}(z_i - \theta_i)^2 + \lambda\sum_{i=1}^{d}|\theta_i|, \tag{2.1.3}$$

where $\lambda = \sigma^2/\lambda_1$.

Minimization of (2.1.3) is equivalent to minimizing (2.1.3) component-wise. The solution to the above problem yields the soft-thresholding rule (Figure 2.1(b))

$$\hat{\theta}_j = \text{sgn}(z_j)(|z_j| - \lambda)_+. \tag{2.1.4}$$

This connection was observed by Donoho, Johnstone, Hoch and Stern (1992) and formed the core of the LASSO method introduced by Tibshirani (1996). If the L_1-penalty in (2.1.3) is replaced by the L_q-penalty for values of $q \geq 1$, it results in the bridge regression proposed by Frank and Friedman (1993) and carefully studied by Fu (1998). Particularly, when $q = 2$, it leads to the usual ridge regression.

2.1.2 Penalized least-squares and variable selection

Consider a general form of penalized least-squares:

$$\frac{1}{2}\sum_{j=1}^{d}(z_j - \theta_j)^2 + \lambda\sum_{j=1}^{d}p_j(|\theta_j|). \tag{2.1.5}$$

The penalty functions $p_j(|\theta|)$ in (2.1.5) are not necessarily the same for all j. For example, one may wish to keep important predictors in a parametric model and hence not be willing to penalize their corresponding parameters. For simplicity of presentation, it is assumed throughout the thesis that the penalty functions for all coefficients are the same, denoted by $p(|\theta|)$.

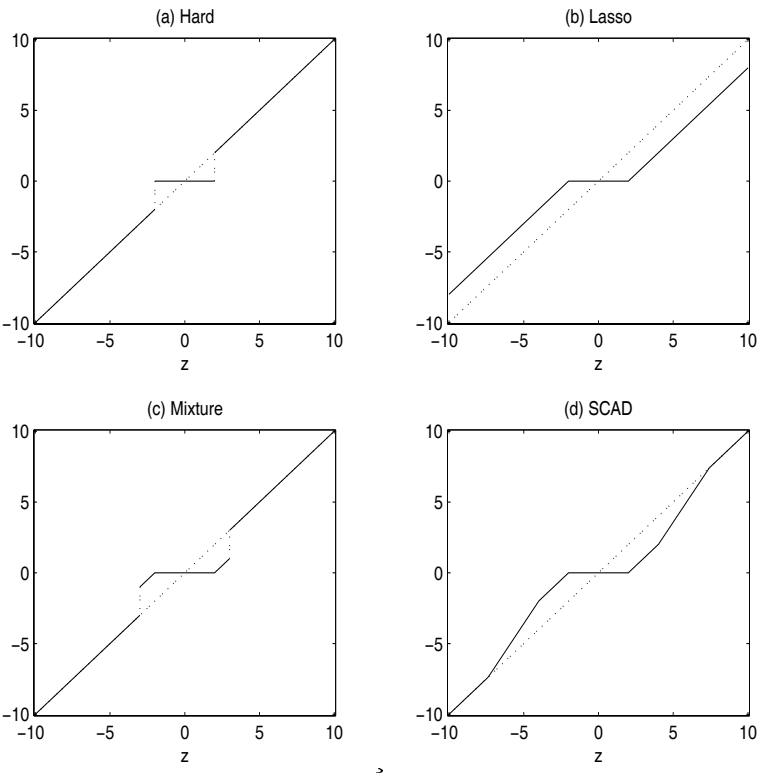


Figure 2.1: *Plot of thresholding functions. (a), (b) and (c) are the hard, soft and mixture thresholding functions with* $\lambda = 2$*, respectively. (d) corresponds to the SCAD (2.1.12) with* $\lambda = 2$ *and* $a = 3.7$.

Furthermore, denote $\lambda p(|\theta|)$ by $p_\lambda(|\theta|)$ so $p(|\theta|)$ can be allowed to depend on λ. Extensions to the case with different penalty functions do not involve any extra difficulties.

A good penalty function should result in an estimator with properties of unbiasedness, sparsity and continuity. In this section we present some mathematical insights and guidance as to how to construct a good penalty function.

The minimization problem of (2.1.5) is equivalent to minimizing componentwise. Thus it is sufficient to consider the penalized least-squares problem:

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|). \tag{2.1.6}$$

The first order derivative of (2.1.6) with respect to θ is $\text{sgn}(\theta)\{|\theta| + p'_\lambda(|\theta|)\} - z$. It is easy to see that when $p'_\lambda(|\theta|) = 0$ for large $|\theta|$, the resulting estimator is unbiased when $|\theta|$ is large. In fact, the condition that $p'_\lambda(|\theta|) = 0$ **for large** $|\theta|$ is a sufficient and necessary condition for unbiasedness for large $|\theta|$. This condition implies that the penalty function $p_\lambda(|\theta|)$ must be a constant for large $|\theta|$. It corresponds to an improper prior distribution in the Bayesian model

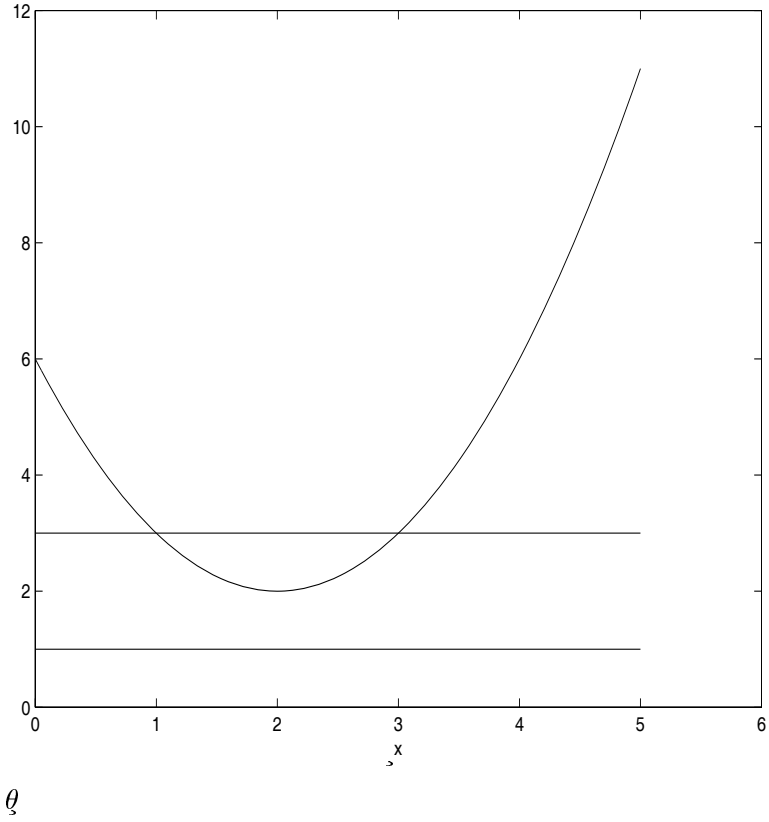


Figure 2.2: *A plot of* $\theta + p'_\lambda(\theta)$ *against* θ *(*$\theta > 0$*).*

selection setting. A sufficient condition for the resulting estimator being a thresholding rule is that **the minimum of the function** $|\theta| + p'_\lambda(|\theta|)$ **is positive**. This condition is referred to as the condition of sparsity. Figure 2.2 provides further insights of this statement. When the values of $|z|$ is less than the minimum of the function $|\theta| + p'_\lambda(|\theta|)$, the derivative of (2.1.6) is positive for all positive θ's, (and is negative for all negative θ's). Therefore the penalized least-squares estimator is 0 in this situation, which implies that the resulting estimator is necessarily a thresholding rule. When the minimum of the function $|\theta| + p'_\lambda(|\theta|)$ is not attained at the origin, and the value of $|z|$ is greater than the minimum. There may exist two crossings as shown in Figure 2.2, and the large one is a penalized least-squares estimator and the small one is not. This implies that a sufficient and necessary condition for continuity is that **the minimum of the function** $|\theta| + p'_\lambda(|\theta|)$ **is attained at 0**. From the above discussion, a penalty function satisfying the conditions of sparsity and continuity must be singular at the origin. For $p_\lambda(|\theta|) = \lambda|\theta|^q$ as in the bridge regression (Frank and Friedman, 1993), the solution is continuous only when $q \geq 1$. However, when $q > 1$, the minimum of $|\theta| + p'_\lambda(|\theta|)$ is zero and hence it does not correspond to a thresholding rule. The only continuous solution with a

thresholding in this family is the L_1 penalty, but this comes at a price of shifting the resulting estimator by a constant λ (see Figure 2.1 (b)).

In the discussion of Antoniadis (1999), Fan observed that the penalized least-squares estimator with the penalty function $p(|\theta|) = |\theta|I(|\theta| \leq \lambda) + \lambda/2I(|\theta| > \lambda)$ leads to the hard-thresholding rule

$$\hat{\theta} = zI(|z| > \lambda). \tag{2.1.7}$$

This penalty function does not over penalize the large value of $|\theta|$. In his response, Antoniadis (1999) improves Fan's proposal by using the following hard thresholding penalty function:

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda). \tag{2.1.8}$$

This is a smoother penalty function than the entropy penalty $p_\lambda(|\theta|) = \frac{\lambda^2}{2} I(|\theta| \neq 0)$, which also results in (2.1.7).

With the clipped L_1-penalty function

$$p_\lambda(|\theta|) = \lambda \min(|\theta|, \lambda) \tag{2.1.9}$$

the solution is a mixture of the soft and hard thresholding rules (Figure 2.1(c)):

$$\hat{\theta} = \text{sgn}(z)(|z| - \lambda)_+ I(|z| \leq 1.5\lambda) + zI(|z| > 1.5\lambda). \tag{2.1.10}$$

2.1.3 Smoothly clipped absolute deviation penalty

From Figure 2.1 (a), (b) and (c), the penalty functions corresponding to the hard, soft and mixture thresholding rules do not simultaneously satisfy the mathematical conditions for unbiasedness, sparsity and continuity. The continuous differentiable penalty function defined by

$$p'_\lambda(\theta) = \lambda\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\} \text{ for some } a > 2 \quad \text{and } \theta > 0, \tag{2.1.11}$$

improves the properties of the L_1-penalty and the hard-thresholding penalty function given by (2.1.8) (see Figure 2.4(d) and discussion below). This penalty function will be called the smoothly clipped absolute deviation (SCAD) penalty. This corresponds to a quadratic spline function with knots at λ and $a\lambda$. This penalty function leaves large values of θ not excessively penalized and makes the solution continuous. The resulting solution is given by

$$\hat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda; \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \leq a\lambda; \\ z, & \text{when } |z| > a\lambda. \end{cases} \tag{2.1.12}$$

See Figure 2.1(d). This solution is due to Fan (1999). For simplicity of presentation, we will use the name SCAD for all procedures using the SCAD penalty. The performance of the SCAD is similar to that of the firm shrinkage proposed by Bruce and Gao (1997) when design matrices are orthonormal. The firm shrinkage is motivated from improving the hard- and soft-thresholding rules. However, the SCAD is motivated from improving their corresponding penalty functions. The view of different shrinkage estimators from a penalized least-squares point of view is part of the contribution of this chapter.

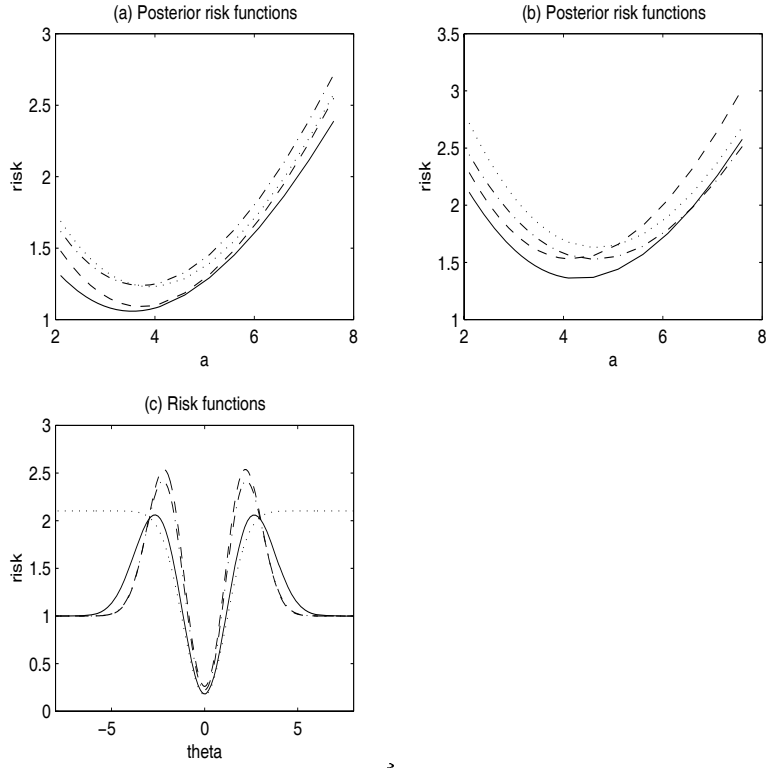


Figure 2.3: *Risk functions of proposed procedures under quadratic loss. (a) and (b) are posterior risk functions of the penalized smoothly clipped-L_1 estimator under the prior $\theta \sim N(0, a\lambda)$ using the universal thresholding $\lambda = \sqrt{2\log(d)}$ for 4 different values of d; the solid, dashed, dashdot and dotted lines are for $d = 20,\ 40,\ 60$ and 100, respectively. (b) is similar to (a) with the solid, dashed, dashdot, dotted lines for $d = 512,\ 1024,\ 2048$ and 4096, separately. (c) Risk functions of the four different thresholding rules. The solid, dashed, dashdot and dotted lines are for minimum SCAD, hard, mixture and soft thresholding rules.*

The thresholding rule in (2.1.12) involves two unknown parameters λ and a. In practice, one could search the best pair (λ, a) over two dimensional grids using some criteria, such as cross-validation and generalized cross-validation (Craven and Wahba, 1977). Such an implementation can be computationally expensive. Motivated by the soft-thresholding, we assume

that for given a and λ, the prior for θ is a normal distribution with zero mean and variance $a\lambda$. We computed the Bayes risk via numerical integration. Figure 2.3(a) depicts the Bayes risk as a function of a under the squared loss, for the universal thresholding $\lambda = \sqrt{2\log(d)}$ (see Donoho and Johnstone, 1994) with $d = 20,\ 40,\ 60$ and 100, and Figure 2.3(b) is for $d = 512$, 1024, 2048 and 4096 . From Figures 2.3(a) and 2.3(b), it can be seen that the Bayes risks are not very sensitive to the value of a. It can be seen from Figure 2.3(a) that the Bayes risks achieve their minimums when $a \approx 3.7$ when the value of d is less than 100. This choice gives pretty good practical performance for various variable selection problems. Indeed, based on the simulations in Section 2.4.3, the choice of $a = 3.7$ works similarly to that chosen by the GCV method.

2.1.4 Performance of thresholding rules

Marron *et al* (1998) applied the tool of exact risk analysis, to understand small sample behavior of wavelet estimators, and to check directly the conclusions suggested by asymptotic analysis. They compared risks of the hard and soft thresholding rules. The closed forms for the L_2 risk functions $R(\theta, \lambda, \sigma) = E(\hat{\theta} - \theta)^2$ have been derived under the Gaussian model $Z \sim N(\theta, \sigma^2)$ for the hard-thresholding and soft-thresholding rules, by Donoho and Johnstone (1994b). It is easy to show that

$$R(\theta, \lambda, \sigma) = \sigma^2 R(\theta/\sigma, \lambda/\sigma, 1)$$

For simplicity, denote by $R(\theta, \lambda)$ the risk function when $\sigma = 1$. Let $\Phi(x)$ denote the standard normal cumulative distribution function and $\phi(x)$ the standard normal density function. When $Z \sim N(\theta, 1)$, for the hard thresholding rule,

$$\begin{aligned} R(\theta, \lambda) &= 1 + (\theta^2 - 1)\{\Phi(\lambda - \theta) - \Phi(-\lambda - \theta)\} \\ &\quad + (\lambda + \theta)\phi(\lambda + \theta) + (\lambda - \theta)\phi(\lambda - \theta), \end{aligned}$$

and for the soft thresholding rule,

$$\begin{aligned} R(\theta, \lambda) &= 1 + \lambda^2 + (\theta^2 - \lambda^2 - 1)\{\Phi(\lambda - \theta) - \Phi(-\lambda - \theta)\} \\ &\quad - (\lambda - \theta)\phi(\lambda + \theta) - (\lambda + \theta)\phi(\lambda - \theta). \end{aligned}$$

For the mixture thresholding and SCAD thresholding rules, risk functions are given in the following theorem.

Theorem 2.1 *When $Z \sim N(\theta, 1)$, for the mixture thresholding rule,*

$$\begin{aligned} R(\theta, \lambda) &= 1 + (\theta^2 - \lambda^2 - 1)\{\Phi(\lambda - \theta) - \Phi(-\lambda - \theta)\} \\ &\quad + \lambda^2\{\Phi(1.5\lambda - \theta) - \Phi(-1.5\lambda - \theta)\} \\ &\quad + 2\lambda\{\phi(1.5\lambda + \theta) + \phi(1.5\lambda - \theta)\} \\ &\quad - (\lambda - \theta)\phi(\lambda + \theta) - (\lambda + \theta)\phi(\lambda - \theta), \end{aligned}$$

and for the SCAD thresholding rule,

$$\begin{aligned} &R(\theta, \lambda) = 1 + (a\lambda + \theta)\phi(a\lambda + \theta) + (a\lambda - \theta)\phi(a\lambda - \theta) \\ &+(\theta - \lambda)\phi(\lambda + \theta) - (\theta + \lambda)\phi(\lambda - \theta) \\ &-\theta\{\phi(2\lambda + \theta) - \phi(2\lambda - \theta)\} + \Phi(-a\lambda - \theta) - \Phi(a\lambda - \theta) \\ &+(1 + \lambda^2 - \theta^2)\{\Phi(-\lambda - \theta) - \Phi(\lambda - \theta)\} \\ &+(1 + \lambda^2)\{\Phi(2\lambda - \theta) - \Phi(-2\lambda - \theta)\} \\ &+\left(\frac{a-1}{a-2}\right)^2\{\Phi(a\lambda - \theta) - \Phi(-a\lambda - \theta) + \Phi(-2\lambda - \theta) - \Phi(2\lambda - \theta)\} \\ &+\left(\frac{a\lambda + \theta}{a-2}\right)^2\{\Phi(-2\lambda - \theta) - \Phi(-a\lambda - \theta)\} \\ &+\left(\frac{a\lambda - \theta}{a-2}\right)^2\{\Phi(a\lambda - \theta) - \Phi(2\lambda - \theta)\} \\ &-\frac{(a-1)(a-3)(a\lambda + \theta)}{(a-2)^2}\phi(a\lambda + \theta) - \frac{(a-1)(a-3)(a\lambda - \theta)}{(a-2)^2}\phi(a\lambda - \theta) \\ &+\frac{(a-1)(a\theta - 2\lambda - 3\theta)}{(a-2)^2}\phi(2\lambda + \theta) - \frac{(a-1)(a\theta + 2\lambda - 3\theta)}{(a-2)^2}\phi(2\lambda - \theta). \end{aligned}$$

The proof of Theorem 2.1 is given in Section 2.5. To gauge the performance of the four thresholding rules, Figure 2.3(c) depicts their L_2 risk functions under the Gaussian model $Z \sim N(\theta, 1)$. To make the scale of the thresholding parameters roughly comparable, we took $\lambda = 2$ for the hard thresholding rule, and adjusted the values of λ for other thresholding rules so that their estimated values are the same when $\theta = 3$. The SCAD performs favorably compared with the other three rules. This can also be understood via their corresponding penalty functions plotted in Figure 2.4. It is clear that the SCAD retains the good mathematical properties of the other three thresholding penalty functions. Hence, it is expected to perform the best. For general σ^2, the picture is the same, except scaled vertically by σ^2, and the θ axis should be replaced by θ/σ.

2.2 Variable Selection via Penalized Likelihood

The methodology in the previous section can be directly applied to many other statistical contexts. This section is devoted to extending the ideas in the last section to general linear regression models, robust linear models and likelihood based generalized linear models. From now on, it is assumed that the design matrix $\mathbf{X} = (x_{ij})$ is standardized so that each column has mean zero and variance one.

2.2.1 Penalized least-squares and likelihood

In classical linear regression models, the least-squares estimate is obtained via minimizing the sum of squared residual errors. Therefore (2.1.5) can be naturally extended to the situation in which design matrices are not orthonormal. Similar to (2.1.3), a general form of penalized least-squares is

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$$

or equivalently

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.2.1}$$

Minimizing (2.2.1) with respect to $\boldsymbol{\beta}$ leads to a penalized least-squares estimator of $\boldsymbol{\beta}$.

It is well known that the least-squares estimate is not robust, one can consider the outlier-resistant loss functions such as the L_1-loss or more generally Huber's ψ-function (see Huber (1981)). Therefore instead of minimizing (2.2.1), we minimize

$$\sum_{i=1}^{n}\psi(|y_i - \mathbf{x}_i\boldsymbol{\beta}|) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|) \tag{2.2.2}$$

with respect to $\boldsymbol{\beta}$. This results in a robust penalized least-squares estimator.

For generalized linear models, statistical inferences are based on underlying likelihood functions. The penalized maximum likelihood estimator can be used to select significant variables. Assume that the collected data $(\mathbf{x}_i, Y_i)$ are independent samples. Conditioning on $\mathbf{x}_i$, Y_i has a density $f_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i)$, where g is a known link function. Let $\ell_i = \log f_i$ denote the conditional log-likelihood of Y_i. A general form of the penalized likelihood is

$$-\sum_{i=1}^{n}\ell_i(g(\mathbf{x}_i^T\boldsymbol{\beta}), y_i) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.2.3}$$

To obtain a penalized maximum likelihood estimator of $\boldsymbol{\beta}$, we minimize (2.2.3) with respect to $\boldsymbol{\beta}$ for some thresholding parameter λ.

2.2.2 A unified algorithm

Finding solutions for minimization problems in (2.2.1), (2.2.2) and (2.2.3) is not an easy task. Tibshirani (1996) proposed an algorithm for solving constrained least-squares problems of LASSO, while Fu (1998) provided a "shooting algorithm" for LASSO. Also see LASSO2 submitted by Berwin Turlach at Statlib (http://lib.stat.cmu.edu/S/). In this section a unified algorithm will be proposed for minimization problems (2.2.1), (2.2.2) and (2.2.3) via local quadratic approximations. The first term in (2.2.1), (2.2.2) and (2.2.3) may be regarded as a loss function of $\boldsymbol{\beta}$. Denote it by $\ell(\boldsymbol{\beta})$. Then, the expressions (2.2.1), (2.2.2) and (2.2.3) can be written in a unified form as

$$\ell(\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{2.2.4}$$

The clipped L_1 penalty $p_\lambda(x)$ in (2.1.9) is not differentiable. However it can be locally approximated by a quadratic function as follows. For given an initial value $\boldsymbol{\beta}_0$ that is close to the minimizer of (2.2.4). Then the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by $\{p_\lambda(|\beta_{j0}|)/\beta_{j0}^2\}\beta_j^2$ for $\beta_j \approx \beta_{j0}$ when β_{j0} is not very close to 0, otherwise, set $\hat{\beta}_j = 0$ (see Figure 2.4(c)). When $p_\lambda(|\beta_j|)$ is differentiable except at the point zero, it can be locally approximated by a quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\mathrm{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}\beta_j,$$

when $\beta_j > 0$. In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}p'_\lambda(|\beta_{j0}|)(\beta_j^2 - \beta_{j0}^2), \qquad \text{for } \beta_j \approx \beta_{j0}.$$

In the algorithm below, the second type of approximation is always applied whenever $p_\lambda(|\beta_j|)$ has the derivative function except at the point 0. Figure 2.4 shows the above two approximations for a few different values of β_{j0}. A drawback of this approximation is that once a coefficient is shrunk to zero, it will stay at zero. However, this method reduces significantly the computational burden.

If $\ell(\boldsymbol{\beta})$ is L_1 loss as used in (2.1.3), then it does not have continuous second order partial derivatives with respect to $\boldsymbol{\beta}$. However, $\psi(|y-\mathbf{x}^T\boldsymbol{\beta}|)$ in (2.2.2) can be analogously approximated by $\{\psi(|y-\mathbf{x}^T\boldsymbol{\beta}_0|)/(y-\mathbf{x}^T\boldsymbol{\beta}_0)^2\}(y-\mathbf{x}^T\boldsymbol{\beta})^2$, as long as the initial value $\boldsymbol{\beta}_0$ of $\boldsymbol{\beta}$ is close to the

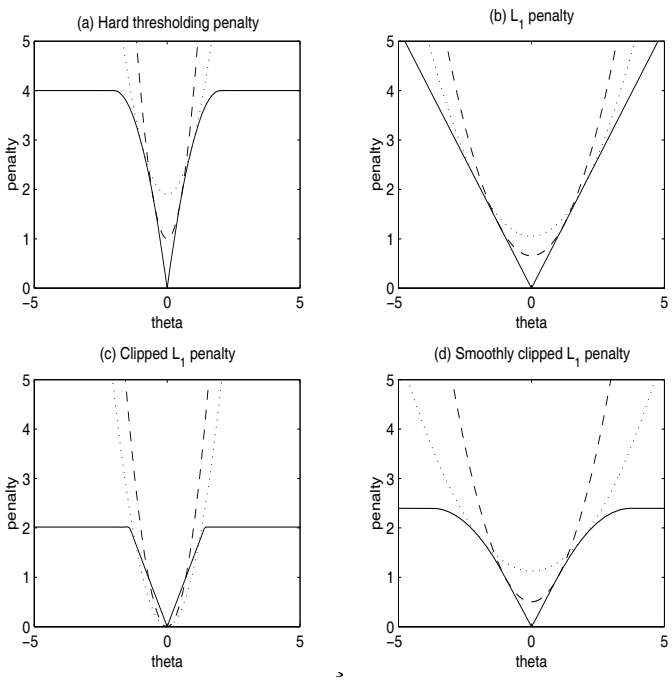


Figure 2.4: *Four penalties $p_\lambda(\theta)$ and their quadratic approximations. The values of λ are the same as those in Figure 2.3(c).*

minimizer. When some of the residuals $|y - \mathbf{x}^T\boldsymbol{\beta}_0|$ are small, this approximation is not very good. See Section 2.2.3 for some slight modification of this approximation.

Now assume that the log-likelihood function is smooth with respect to $\boldsymbol{\beta}$ so that its first two partial derivatives are continuous. Thus the first terms in (2.2.4) can be locally approximated by a quadratic function. Therefore the minimization problem (2.2.4) can be reduced to a quadratic minimization problem and the Newton-Raphson algorithm can be used. In particularly, when $p'_\lambda(|\beta|)$ has the first derivative except at the point 0, (2.2.4) can be locally approximated (except for a constant term) by

$$\ell(\boldsymbol{\beta}_0) + \nabla\ell(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\nabla^2\ell(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \tag{2.2.5}$$

where

$$\nabla\ell(\boldsymbol{\beta}_0) = \frac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}, \quad \nabla^2\ell(\boldsymbol{\beta}_0) = \frac{\partial^2\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}, \quad \Sigma_\lambda(\boldsymbol{\beta}_0) = \mathrm{diag}\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \cdots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}.$$

The quadratic minimization problem (2.2.5) yields the solution

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \{\nabla^2\ell(\boldsymbol{\beta}_0) + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\{\nabla\ell(\boldsymbol{\beta}_0) + n\mathbf{U}_\lambda(\boldsymbol{\beta}_0)\}, \tag{2.2.6}$$

where $\mathbf{U}_\lambda(\boldsymbol{\beta}_0) = \Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0$. When the algorithm converges, and the second type of approximation is used, the estimator satisfies the condition

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

the penalized likelihood equation, for non-zero elements of $\hat{\boldsymbol{\beta}}_0$. Specifically, for the penalized least-squares problem (2.1.2), the solution can be found by iteratively computing the following ridge regression:

$$\boldsymbol{\beta}_1 = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}.$$

Similarly the solution for (2.1.3) can be obtained by iterating

$$\boldsymbol{\beta}_1 = \{\mathbf{X}^T\mathbf{W}\mathbf{X} + \frac{1}{2}n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.$$

where $\mathbf{W} = \mathrm{diag}\{\psi(|y_1 - \mathbf{x}_1^T\boldsymbol{\beta}_0|)/(y_1 - \mathbf{x}_1^T\boldsymbol{\beta}_0)^2, \cdots, \psi(|y_n - \mathbf{x}_n^T\boldsymbol{\beta}_0|)/(y_n - \mathbf{x}_n^T\boldsymbol{\beta}_0)^2\}$.

As in the maximum likelihood estimation (MLE) setting, with good initial value $\boldsymbol{\beta}_0$, the one-step procedure can be as efficient as the fully iterative procedure, namely, the penalized maximum likelihood estimator, when one uses the Newton-Raphson algorithm (See Bickel (1975)). Now regarding $\boldsymbol{\beta}^{(k-1)}$ as a good initial value at the k-th step, the next iteration can also be regarded as a one-step procedure and hence the resulting estimator can still be as efficient as the fully iterative method. See Robinson (1988) for theory on the difference between the MLE and k-step estimators. Therefore estimators obtained by the aforementioned algorithm after a few iterations can always be regarded as a one-step estimator, which is as efficient as the fully iterative method. In this sense, one does not have to iterate the algorithm above until it converges as long as the initial estimators are good enough. The estimators from the full models can be used as initial estimators, as long as they are not excessively overly parameterized.

2.2.3 Standard error formula

The standard errors for estimated parameters can be directly obtained because we are estimating parameters and selecting variables at the same time. Following the conventional technique in the likelihood setting, the corresponding sandwich formula in (2.2.6) can be used as an estimator for the conditional covariance of the estimates $\hat{\boldsymbol{\beta}}$, conditioning on $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$. That is,

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = \{\nabla^2\ell(\hat{\boldsymbol{\beta}}) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}\widehat{\mathrm{cov}}\{\nabla\ell(\hat{\boldsymbol{\beta}})\}\{\nabla^2\ell(\hat{\boldsymbol{\beta}}) + n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}. \tag{2.2.7}$$

This formula will be shown to have good accuracy for moderate sample sizes.

When the L_1-loss is used in the robust regression, some slight modifications are needed in the aforementioned algorithm and its corresponding sandwich formula. For $\psi(x) = |x|$, the diagonal elements of $\mathbf{W}$ are $\{|r_i|^{-1}\}$ with $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0$. Thus, for a given current value of $\boldsymbol{\beta}_0$, when some of the residuals $\{r_i\}$ are close to 0, these points receive too much weights. Hence in practical implementation one may replace the weight by $(a_n + |r_i|)^{-1}$ with a_n being the $2n^{-1/2}$ quantile of the absolute residuals $\{|r_i|, i = 1, \cdots, n\}$. Thus, the constant a_n is changing from iteration to iteration.

2.2.4 Testing convergence of the algorithm

To demonstrate that the proposed algorithm converges to the right solution, the following small experiment was conducted: first take a 100 dimensional vector $\boldsymbol{\beta}$ consisting of 50 zeros and other nonzero elements being generated from $N(0, 5^2)$, and use a 100×100 orthonormal design matrix $\mathbf{X}$. Then generate a response vector $\mathbf{y}$ from the linear model (2.1.1). The reason for choosing an orthonormal design matrix for the testing case is that the penalized least-squares has a closed form mathematical solution so that comparison outputs with the mathematical solution can be conducted. This experiment did show that the proposed algorithm converged to the right solution. It took MATLAB 0.27, 0.39, 0.16 seconds for the penalized least-squares with the SCAD, L_1 and hard-thresholding penalties to converge. The numbers of iterations are respectively 30, 30 and 5 for the penalized least-squares with the SCAD, L_1 and the hard-thresholding penalty. In fact, after 10 iterations, the penalized least-squares estimators are already very close to the true one.

2.3 Examples

In this section we illustrate our approaches via application of two real data examples. We first apply the proposed penalized least-squares approaches to a *Female Labor Supply* data set from East Germany collected around 1994 (see Fan, Härdle and Mammen, 1998). We also apply the proposed penalized likelihood method to analyze another data set: the *Burns data*, collected by the General Hospital Burn Center at the University of Southern California. We computed the penalized likelihood estimate with L_1 penalty, referred to as LASSO, by our algorithm rather than those of Tibshirani (1996) and Fu (1998). The purpose of this section is

to illustrate our approaches via real data examples and to compare them with the best subset variable selection. In the following two examples, they do not involve new regression techniques to solve.

Example 2.3.1. We now illustrate the proposed penalized least-squares approaches by the "*Female Labor Supply*" data. This data set was analyzed, using an additive partial linear model, by Fan, Härdle and Mammen (1998). They used local polynomial regression techniques to estimate the nonlinear components. In this example, we parameterize some component functions by linear spline basis functions. A brief description of this data set is given below, see Fan *et al* (1998) for details. There are 607 observations in this data set. Take the weekly number Y of working hours as the response variable. Seven covariates are considered: if the woman has children less than 16 years old (X_1), the unemployment rate (X_2) in the place where she lives, the age (X_3) of the woman, her years (X_4) of education, the "Treiman prestige index" of her job (X_5), her wage per hour (X_6) and the monthly net income (X_7) of her husband. Based on Figure 1 of Fan *et al* (1998), we fit the data by the model

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_5 X_5 + \beta_6 X_3^2 + f_1(X_6) + f_2(X_7) + X_3 f_3(X_6),$$

and parameterize f_j by linear spline functions with knots at the quartiles of variables. Specifically, $f_1(X_6) = \beta_{60} X_6 + \beta_{61}(X_6 - k_1)_+ + \beta_{62}(X_6 - k_2)_+ + \beta_{63}(X_6 - k_3)_+$, where k_i's are the 1st, 2nd and 3rd quartiles of X_6. Denote by X_{60}, X_{61}, X_{62}, X_{63} the variables X_6, $(X_6 - k_1)_+$, $(X_6 - k_2)_+$, $(X_6 - k_3)_+$, respectively. Similar notation applies to $f_2(X_7)$ and $f_3(X_6)$. This gives a total of 18 predictors. Before using the penalized least-squares method, the response variable Y and all 18 covariates were standardized individually. We fitted a linear regression model without intercept as both the response and the predictors were normalized. The thresholding parameter λ is chosen by generalized cross-validation (see Section 2.4.2 below). They are 0.0301, 0.0007 and 0.0188 for the penalized least-squares with the SCAD, L_1 (LASSO) and hard thresholding penalties, respectively. The value of a in the SCAD was set to be 3.7. To find the best subset, we searched exhaustively over all possible subsets and selected the subset with the best BIC score. All computations were conducted on a Pentium II 450 PC with MATLAB code. The computational time for each of the three penalized least-squares including searching for the unknown parameter λ over 15 grids via generalized cross-validation was about one minute; while it spent about half an hour in searching for the best subset using the naive implementation. With the selected λ, the penalized least-squares estimators were obtained at the 9^{th}, 6^{th} and 3^{rd} step of the iterations for the SCAD, LASSO and hard thresh-

olding, respectively. Estimated coefficients, standard errors, AIC, BIC and Mallow's C_p scores for the transformed data are presented in Table 2.1.

Table 2.1: Estimated coefficients and standard errors for Example 2.3.1

Method	LS	best subset	SCAD	LASSO	hard
X_1	-0.15(0.06)	0(−)	-0.14(0.06)	-0.14(0.06)	-0.14(0.06)
X_2	0.08(0.04)	0(−)	0.05(0.03)	0.08(0.04)	0.08(0.04)
X_3	1.53(0.40)	1.47(0.40)	1.49(0.40)	1.31(0.35)	1.48(0.40)
X_4	0.19(0.05)	0.18(0.05)	0.18(0.05)	0.19(0.05)	0.18(0.05)
X_5	0.23(0.07)	0.27(0.05)	0.23(0.05)	0.24(0.05)	0.23(0.05)
X_3^2	-1.50(0.40)	-1.41(0.40)	-1.52(0.40)	-1.34(0.36)	-1.51(0.40)
X_{60}	-1.17(0.30)	-0.29(0.05)	-1.03(0.28)	-0.94(0.26)	-1.04(0.28)
X_{61}	-0.15(3.00)	0(−)	0(−)	0(−)	0(−)
X_{62}	1.33(0.50)	0(−)	0.75(0.28)	0.66(0.26)	0.76(0.28)
X_{63}	-0.25(2.86)	0(−)	0(−)	0(−)	0(−)
X_{70}	0.00(0.15)	0(−)	0(−)	0(−)	0(−)
X_{71}	-1.01(0.51)	0(−)	-0.90(0.38)	-0.76(0.32)	-0.90(0.38)
X_{72}	0.92(0.43)	0(−)	0.85(0.36)	0.72(0.30)	0.86(0.36)
X_{73}	0.08(0.18)	0(−)	0(−)	0(−)	0(−)
$X_3 * X_{60}$	-0.18(0.15)	-0.22(0.05)	-0.14(0.07)	-0.14(0.07)	-0.14(0.07)
$X_3 * X_{61}$	0.36(2.89)	0(−)	0(−)	0(−)	0(−)
$X_3 * X_{62}$	-0.36(0.39)	0(−)	0(−)	0(−)	0(−)
$X_3 * X_{63}$	-0.07(2.76)	0(−)	0(−)	0(−)	0(−)
BIC		-47.2099	-35.1394	-35.0377	-35.4736
AIC		-73.6610	-83.6333	-83.5315	-83.9674
C_p		19.8767	9.9745	10.0743	9.6470

From Table 2.1, it can be seen that the three penalized least-squares approaches choose the same covariates. Best subset selection deletes a lot of covariates, compared with the penalized least-squares approaches. Models selected via the penalized least-squares approaches are consistent with the partial linear model considered by Fan *et al* (1998). The number of working hours depends on the variables of children, unemployment rate, age, years of education, job prestige linearly, and hourly income and monthly net income of her husband nonlinearly. It may be concluded that there is interaction between the variables *age* and *hourly income* in the form of X_3X_6.

By our construction, the best subset selection has the least BIC score. However, it has the greatest AIC and Mallow's C_p scores. The three penalized least-squares methods yield similar models. Their AIC, BIC and C_p scores are almost the same. In this example, the sample size n is 607, which is quite large, and hence the asymptotic properties in the next chapter

are expected to hold. Therefore the penalized least-squared approaches are expected to select the correct model. The SCAD and the penalized least-squares with the hard-thresholding penalty result in two unbiased estimators for unknown coefficient. While LASSO results in a biased estimator for large coefficients. This can be seen from Table 2.1 by comparing the large estimated coefficients of LASSO with the ordinary least-squares estimates.

Example 2.3.2 In this example the proposed penalized likelihood methodology is applied to the *Burns data*. The data set consists of 981 observations. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise. Covariates $X_1 = age$, $X_2 = sex$, X_3 =log(burn area +1) and binary variable $X_4 = Oxygen$ (0 normal, 1 abnormal) were considered. Quadratic terms of X_1 and X_3, and all interaction terms were included. The intercept term was added and the logistic regression model was fitted. The unknown parameter λ was chosen by generalized cross-validation. They are 0.6932, 0.0015 and 0.8062 for the penalized likelihood estimates with the SCAD, L_1 and hard-thresholding penalties respectively. The constant a in the SCAD was taken as 3.7. With the selected λ, the penalized likelihood estimator was obtained at the 6^{th}, 28^{th} and 5^{th} step iterations for the penalized likelihood with the SCAD, L_1 and hard-thresholding penalties, respectively. Ten-step estimators were also computed, it took us less than 50 seconds for each penalized likelihood estimator, and the differences between the full iteration estimators and the ten-step estimators were less than one percent. The estimated coefficients and standard errors for the transformed data, based on the penalized likelihood estimators, are reported in Table 2.2.

From Table 2.2, the best subset procedure chooses 5 out of 13 covariates, while the SCAD chooses 4 covariates. The difference between them is that the best subset keeps X_4. LASSO chooses the quadratic term of X_1 and X_3 rather than their linear terms. It also selects an interaction term X_2X_3, which may not be statistically significant. It seems again that LASSO shrinks noticeably large coefficients. In this example, the penalized likelihood with the hard thresholding penalty retains too many predictors. Particularly, it selects variables X_2 and X_2X_3.

Table 2.2: Estimated coefficients and standard errors for Example 2.3.2

Method	MLE	best subset	SCAD	LASSO	hard
intercept	5.51 (0.75)	6.12 (0.57)	6.09 (0.29)	3.70 (0.25)	5.88 (0.41)
X_1	-8.83 (2.97)	-12.15 (1.81)	-12.24 (0.08)	0 (−)	-11.32 (1.1)
X_2	2.30 (2.00)	0 (−)	0 (−)	0 (−)	2.21 (1.41)
X_3	-2.77 (3.43)	-6.93 (0.79)	-7.00 (0.21)	0 (−)	-4.23 (0.64)
X_4	-1.74 (1.41)	-0.29 (0.11)	0 (−)	-0.28 (0.09)	-1.16 (1.04)
X_1^2	-0.75 (0.61)	0 (−)	0 (−)	-1.71 (0.24)	0 (−)
X_3^2	-2.70 (2.45)	0 (−)	0 (−)	-2.67 (0.22)	-1.92 (0.95)
X_1X_2	0.03 (0.34)	0 (−)	0 (−)	0 (−)	0 (−)
X_1X_3	7.46 (2.34)	9.83 (1.63)	9.84 (0.14)	0.36 (0.22)	9.06 (0.96)
X_1X_4	0.24 (0.32)	0 (−)	0 (−)	0 (−)	0 (−)
X_2X_3	-2.15 (1.61)	0 (−)	0 (−)	-0.10 (0.10)	-2.13 (1.27)
X_2X_4	-0.12 (0.16)	0 (−)	0 (−)	0 (−)	0 (−)
X_3X_4	1.23 (1.21)	0 (−)	0 (−)	0 (−)	0.82 (1.01)

2.4 Simulations

The purpose of this section is to compare the performance of the proposed approaches and existing ones and to test the accuracy of the standard error formula.

2.4.1 Prediction and model error

The prediction error is defined as the average error in prediction of Y given $\mathbf{x}$ for future cases not used in the construction of a prediction equation. There are two regression situations, *X-random* and *X-controlled*. In the case that X is random, both Y and $\mathbf{x}$ are randomly selected. In the controlled situation, design matrices are selected by experimenters and only y is random. For ease of presentation, we consider only the X-random case.

In *X-random* situations, the data $(\mathbf{x}_i, Y_i)$ are assumed to be a random sample from their parent distribution $(\mathbf{x}, Y)$. Then, if $\hat{\mu}(\mathbf{x})$ is a prediction procedure constructed using the present data, the prediction error is defined as

$$\mathrm{PE}(\hat{\mu}) = E\{Y - \hat{\mu}(\mathbf{x})\}^2,$$

where the expectation is only taken with respect to the new observation $(\mathbf{x}, Y)$. The prediction error can be decomposed as

$$\mathrm{PE}(\hat{\mu}) = E\{Y - E(Y|\mathbf{x})\}^2 + E\{E(Y|\mathbf{x}) - \hat{\mu}(\mathbf{x})\}^2.$$

The first component is inherent by due to stochastic errors. The second component is due to lack of fit to an underlying model. This component is called *model error* and is denoted by $\mathrm{ME}(\hat{\mu})$. The size of the model error reflects performances of different model selection procedures. If $Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon$, where $E(\varepsilon|\mathbf{x}) = 0$, then $\mathrm{ME}(\hat{\mu}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

2.4.2 Selection of thresholding parameters

To implement the methods described in Sections 2 and 3, one has to estimate the thresholding parameters λ and a. Denote by $\boldsymbol{\theta}$ the tuning parameters to be estimated, i.e., $\boldsymbol{\theta} = (\lambda, a)$ for the SCAD, while $\boldsymbol{\theta} = \lambda$ for the other thresholdings. As suggested by Breiman (1995), Tibshirani (1996) and Fu (1998), fivefold cross-validation and generalized cross-validation will be used to estimate $\boldsymbol{\theta}$:

For completeness, here is a brief description of the cross-validation procedures. Let $\ell\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ denote the first term in (2.2.4) replacing $\boldsymbol{\beta}$ by its estimate $\hat{\boldsymbol{\beta}}$ obtained when the tuning parameters $\boldsymbol{\theta}$ are used. Then $\ell\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ can be regarded as a measure of goodness of fit. The fivefold cross-validation procedure is as follows: Denote the full training set by T, and cross-validation training and test set by $T - T^\nu$ and T^ν, for $\nu = 1, \cdots, 5$. For each $\boldsymbol{\theta}$ and ν, find the estimator $\hat{\boldsymbol{\beta}}^{(\nu)}(\boldsymbol{\theta})$ of $\boldsymbol{\beta}$ using the training set $T - T^\nu$. Let $\ell_\nu\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ be the $\ell\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}$ for the test set T^ν. Form the cross-validation criterion as

$$\mathrm{CV}(\boldsymbol{\theta}) = \sum_{\nu=1}^{5} \ell_\nu\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}.$$

Find a $\hat{\boldsymbol{\theta}}$ that minimizes $\mathrm{CV}(\boldsymbol{\theta})$.

The second method is generalized cross-validation. For linear regression models, one updates the solution by

$$\boldsymbol{\beta}_1(\boldsymbol{\theta}) = \{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}.$$

Thus the fitted value $\hat{\mathbf{y}}$ of $\mathbf{y}$ is $\mathbf{X}^T\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}$, and

$$\mathbf{P}_\mathbf{X}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\} = \mathbf{X}^T\{\mathbf{X}^T\mathbf{X} + n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}\mathbf{X}^T$$

can be regarded as a projection matrix. Define the number of effective parameters in the penalized least-squares fit as $e(\boldsymbol{\theta}) = \mathrm{tr}[\mathbf{P}_\mathbf{X}\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}]$. Therefore the generalized cross-validation statistic is

$$\mathrm{GCV}(\boldsymbol{\theta}) = \frac{1}{n}\frac{\ell\{\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}}{\{1 - e(\boldsymbol{\theta})/n\}^2}$$

and $\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\{\text{GCV}(\boldsymbol{\theta})\}$. Similarly the corresponding generalized cross-validation statistics can be defined for robust regression models and likelihood based linear models.

2.4.3 Simulation study

Numerical comparison of the proposed variable selection methods with ordinary least-squares, ridge regression, best subset selection and non-negative garrote (see Breiman (1995)) are given in the following examples. All simulations are conducted using MATLAB codes. The constraint least-squares module in MATLAB was directly utilized for finding the non-negative garrote estimate. As recommended in Breiman (1995), a five-fold cross-validation was used to estimate the tuning parameter for the non-negative garrote. For other model selection procedures, both five-fold cross-validation and generalized cross-validation were used for estimating thresholding parameters. However, their performance was similar. Therefore only the results based on the generalized cross validation are presented.

Example 2.4.1.(Linear regression)

This example simulated 100 data sets consisting of n observations from the model (Tibshirani, 1996)

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \sigma\varepsilon,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and the components of $\mathbf{x}$ and ε are standard normal. The correlation between x_i and x_j is $\rho^{|i-j|}$ with $\rho = 0.5$. This is a model used in Tibshirani (1996). Firstly we chose $n = 40$ and $\sigma = 3$, then reduced σ to 1 and finally increased the sample size to 60. The model error of the proposed procedures are compared to that of the least-squares estimator. The Median of Relative Model Errors (MRME) over 100 simulated data sets are summarized in Table 2.3. The top panel of Figure 2.5 depicts the boxplots of the relative model errors. The average of 0 coefficients is also reported in Table 2.3, in which the column labeled "correct" presents the average restricted only to the true zero coefficients, while the column labeled "incorrect" depicts the average of coefficients erroneously set to 0.

From Figure 2.5 and Table 2.3, it can be seen that when the noise level is high and sample size is small, LASSO performs the best and it significantly reduces both model error and model complexity; while ridge regression only reduces model error. The other variable se-

Table 2.3: Simulation results for linear regression models

Method	MRME(%)	Aver. no. of 0 Coeff.	
$n = 40,\ \sigma = 3$		correct	incorrect
SCAD^1	72.90	4.20	0.21
SCAD^2	69.03	4.31	0.27
LASSO	63.19	3.53	0.07
Hard	73.82	4.09	0.19
Ridge	83.28	0	0
Best subset	68.26	4.50	0.35
Garrote	76.90	2.80	0.09
$n = 40,\ \sigma = 1$		correct	incorrect
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
$n = 60,\ \sigma = 1$		correct	incorrect
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

Note that the value of a SCAD^1 is obtained by generalized cross-validation, while the value of a in SCAD^2 is 3.7.

Table 2.4: Standard deviations of estimators for linear regression models ($n = 60$)

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD^1	0.166	0.161 (0.021)	0.170	0.160 (0.024)	0.148	0.145 (0.022)
SCAD^2	0.161	0.161 (0.021)	0.164	0.161 (0.024)	0.151	0.143 (0.023)
LASSO	0.164	0.154 (0.019)	0.173	0.150 (0.022)	0.153	0.142 (0.021)
Hard	0.169	0.161 (0.022)	0.174	0.162 (0.025)	0.178	0.148 (0.021)
Best subset	0.163	0.155 (0.020)	0.152	0.154 (0.026)	0.152	0.139 (0.020)

lection procedures also reduce model error and model complexity. However, when the noise level is reduced, the SCAD outperforms the LASSO and other penalized least-squares. Ridge regression performs very poorly. The best subset selection method performs quite similarly to the SCAD. The nonnegative garrote performs quite well in various situations. Comparing the first two rows in Table 2.3, one can see that the choice of $a = 3.7$ is very reasonable. Therefore it will be used for other examples in this paper.

Now the accuracy of our standard error formula (2.2.7) is investigated. The median absolute deviation divided by 0.6745, denoted by SD in Table 2.4, of 100 estimated coefficients in the 100 simulations can be regarded as the true standard error. The median of the 100 estimated SDs, denoted by SD_m, and the median absolute deviation error of the 100 estimated standard errors divided by 0.6745, denoted by SD_{mad}, gauge the overall performance of the standard error formula (2.2.7). Table 2.4 presents only the results for non-zero coefficients when the sample size $n = 60$. The results for the other two cases with $n = 40$ are similar. Table 2.4 suggests that the sandwich formula performs surprisingly well.

Example 2.4.2. (Robust regression)

This example simulated 100 data sets consisting of 60 observations from the model

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta}$ and $\mathbf{x}$ are the same as those in Example 1. The ε is drawn from the standard normal distribution with 10% outliers from the standard Cauchy distribution. The simulation results are summarized in Table 2.5. Figure 2.5(d) presents the boxplots of the relative model errors. From Table 2.5, it can be seen that the SCAD outperforms somewhat the other procedures. The true and estimated standard deviations of estimators via sandwich formula (2.1.8) are shown in Table 2.6. It indicates that the performance of the sandwich formula is very good.

Example 2.4.3. (Logistic regression)

This example simulated 100 data sets consisting of 200 observations from the model $Y \sim$ Bernoulli$(p(\mathbf{x}^T\boldsymbol{\beta}))$, where $p(u) = \exp(u)/(1 + \exp(u))$, and the first six components of $\mathbf{x}$ and $\boldsymbol{\beta}$ are the same as those in Example 1. The last two components of $\mathbf{x}$ are independently identically distributed as a Bernoulli distribution with probability of success 0.5. All covariates

Table 2.5: Simulation results for robust linear models

Method	MRME(%)	Aver. no. of 0 Coeff.	
		correct	incorrect
SCAD (a=3.7)	35.52	4.71	0
LASSO	52.80	4.29	0
Hard	47.22	4.70	0
Best subset	41.53	4.85	0.18

Table 2.6: Standard deviations of estimators for robust regression models

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.167	0.171 (0.018)	0.185	0.176 (0.022)	0.165	0.155 (0.020)
LASSO	0.158	0.165 (0.022)	0.159	0.167 (0.020	0.182	0.154 (0.019)
Hard	0.179	0.168 (0.018)	0.176	0.176 (0.025)	0.157	0.154 (0.02)
Best subset	0.198	0.172 (0.023)	0.185	0.175 (0.024)	0.199	0.152 (0.023)

are standardized. Model errors are computed via 1000 Monte Carlo simulations. The summary of simulation results is depicted in Tables 2.7 and 2.8. Figure 2.5(e) shows the boxplots of the relative model errors. From Table 2.7, it can be seen that the performance of the SCAD is much better than the other two penalized likelihood estimates. From Figure 2.5(e), the variations of the relative model errors of the four procedures are almost the same. It can be seen from Table 2.8 that our standard error estimator works well.

Table 2.7: Simulation results for logistic regression

Method	MRME(%)	Aver. no. of 0 Coeff.	
		correct	incorrect
SCAD(a=3.7)	26.48	4.98	0.04
LASSO	53.14	3.76	0
Hard	59.06	4.27	0
Best subset	31.63	4.84	0.01

It is worthy to remark that the estimated SDs for L_1-penalized likelihood estimator (LASSO) are consistently smaller than the SCAD and the penalized likelihood method with the hard-thresholding procedure, yet its overall MRME is larger than that of the SCAD. This implies that the biases in the L_1-penalized likelihood estimators are large. This remark applies to all

Table 2.8: Standard deviations of estimators for logistic regression

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD ($a = 3.7$)	0.571	0.538 (0.107)	0.383	0.372 (0.061)	0.432	0.498 (0.065)
LASSO	0.310	0.379 (0.037)	0.285	0.284 (0.019)	0.244	0.287 (0.019)
Hard	0.675	0.561 (0.126)	0.428	0.400 (0.062)	0.467	0.421 (0.079)
Best subset	0.624	0.547 (0.121)	0.398	0.383 (0.067)	0.468	0.412 (0.077)

of our examples. Indeed, it can be seen from Table 2.8 that all coefficients were noticeably shrunken by LASSO.

Example 2.4.4. (Poisson log-linear regression)

This example simulated 100 data sets consisting of 60 observations from the model $Y \sim \text{Poisson}\{\lambda(\mathbf{x}^T\boldsymbol{\beta})\}$, where $\lambda(u) = \exp(u)$, $\mathbf{x}$ is the same as that in Example 5.3, and $\boldsymbol{\beta} = (1.2, 0.6, 0, 0, 0.8, 0, 0, 0)^T$. Model errors were obtained by 1000 Monte Carlo simulations. Tables 2.9 and 2.10 show the simulation results. Figure 2.5(f) depicts the boxplots of the relative model errors. From Figure 2.5(f), the variations of the relative model errors of the four procedures are almost same. In terms of model errors, the performance of the best subset selection method and the SCAD are much better than the other two. Table 2.10 shows that the standard error estimator works very reasonably.

Table 2.9: Simulation results for Poisson log-linear regression

Method	MRME(%)	Aver. no. of 0 Coeff.	
		correct	incorrect
SCAD(a=3.7)	48.00	3.60	0.01
LASSO	60.93	3.55	0.01
Hard	70.07	3.65	0.01
Best subset	33.96	4.70	0.01

2.5 Proofs

Proof of Theorem 2.1

Table 2.10: Standard deviations of estimators for Poisson log-linear models

	$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_5$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD ($a = 3.7$)	0.080	0.079 (0.014)	0.093	0.084 (0.016)	0.079	0.072 (0.016)
LASSO	0.086	0.078 (0.013)	0.101	0.082 (0.016)	0.083	0.074 (0.017)
Hard	0.084	0.080 (0.015)	0.100	0.086 (0.019)	0.081	0.075 (0.020)
Best subset	0.081	0.079 (0.016)	0.080	0.083 (0.018)	0.079	0.068 (0.016)

By integration by parts, it can be shown that

$$\int_a^b z\phi(z)\,dz = \phi(a) - \phi(b), \tag{2.5.1}$$

$$\int_a^b z^2\phi(z)\,dz = a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a). \tag{2.5.2}$$

These two formulas are useful in the proof of Theorem 2.1. Since the mixture thresholding rule is

$$\hat{\theta}_m = \mathrm{sgn}(z)(|z| - \lambda)_+ I(|z| \le 1.5\lambda) + zI(|z| > 1.5\lambda),$$

when $Z \sim N(\theta, 1)$,

$$\begin{aligned}
& E(\hat{\theta}_m - \theta)^2 \\
= \; & 1 - \int_{-1.5\lambda}^{1.5\lambda} (z-\theta)^2\phi(z-\theta)\,dz \\
& + \int_{-1.5\lambda}^{-\lambda} (z-\theta+\lambda)^2\phi(z-\theta)\,dz + \int_{-\lambda}^{\lambda} \theta^2\phi(z-\theta)\,dz \\
& + \int_{\lambda}^{1.5\lambda} (z-\theta-\lambda)^2\phi(z-\theta)\,dz \\
= \; & 1 - \int_{-1.5\lambda-\theta}^{1.5\lambda-\theta} z^2\phi(z)\,dz \\
& + \int_{-1.5\lambda-\theta}^{-\lambda-\theta} (z+\lambda)^2\phi(z)\,dz + \int_{-\lambda-\theta}^{\lambda-\theta} \theta^2\phi(z)\,dz \\
& + \int_{\lambda-\theta}^{1.5\lambda-\theta} (z-\lambda)^2\phi(z)\,dz
\end{aligned}$$

Using (2.5.1) and (2.5.2) and straightforward algebraic calculation, the expression of risk function for the mixture thresholding rule can be obtained.

The derivation of risk function for SCAD thresholding rule is similar to that for the mixture thresholding rule, but it involves more complicated algebraic calculations. I omit the details here.

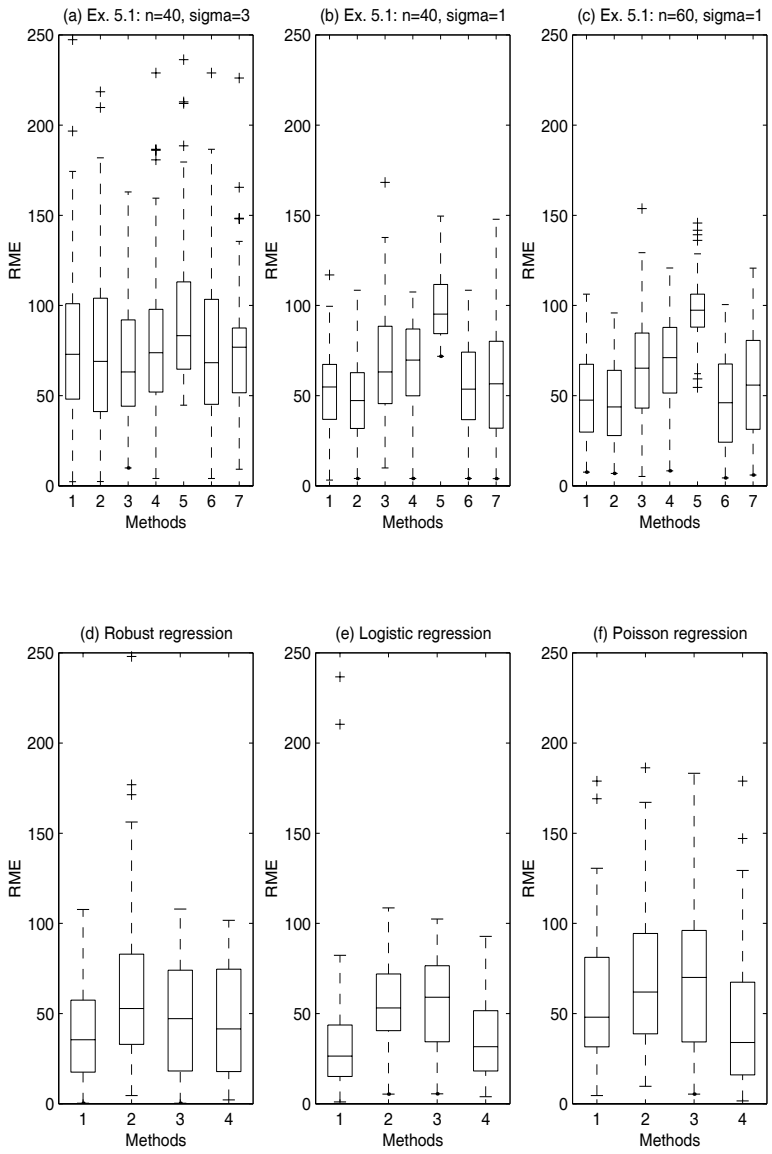


Figure 2.5: *Boxplots of relative model errors. From left to right, the order in the top panel is $SCAD^1$, $SCAD^2$, LASSO, hard, ridge, best subset and nonnegative garrote. The order from left to right in the bottom panel is the SCAD (a=3.7), LASSO, hard and best subset.*

2.6 Conclusion

In this chapter, a few variable selection methods were proposed via penalized likelihood approaches. A family of penalty functions were introduced. The methods were shown to be effective and the standard errors were estimated with good accuracy. A unified algorithm was proposed for minimizing the penalized likelihood function, which is usually a sum of convex

and concave functions. The proposed algorithm was backed up by statistical theory and hence gave estimators with good statistical properties. Comparing with the best subset method, which is very time consuming, the newly proposed methods are much faster, more effective and have strong theoretical backup. They select variables simultaneously via optimizing a penalized likelihood and hence the standard errors of estimated parameters can be estimated accurately. The LASSO proposed by Tibshirani (1996) is a member of this penalized likelihood family with L_1-penalty. It has good performance when noise to signal ratio is large, but the bias created by this approach is noticeably large. See also the remarks in Example 2.4.3. The newly proposed penalty function, called the Smoothly Clipped Absolute Deviation (SCAD) penalty function, gave the best performance in selecting significant variables without creating excessive biases. The approach proposed here can be applied to other statistical contexts without any extra difficulties.

Chapter 3

Theoretical Results and Oracle Properties

3.1 Introduction

This chapter will illustrate that the proposed estimators in Chapter 2 perform as well as an oracle estimator, in the terminology of Donoho and Johnstone (1994a). The oracle performance is closely related to the super-efficiency phenomenon. Consider the simplest linear regression model

$$\mathbf{y} = \mathbf{1}_n \mu + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, I_n)$. A superefficient estimate for μ is

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

due to Hodges, see for example page 405 of Lehmann (1983). If we take c to 0, then δ_n coincides with the hard-thresholding estimator with the thresholding parameter $\lambda_n = n^{-1/4}$. This estimator correctly estimates the parameter at the point 0 without paying an asymptotic price in terms of estimating the parameter elsewhere.

Let the true value of $\boldsymbol{\beta}$ be

$$\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T,$$

which is a fixed constant. Without loss of generality, assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$. It will be shown

that the resulting estimators in Chapter 2 converge at the rate

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

where $a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$. This implies that for the hard thresholding and SCAD penalty functions, the penalized likelihood estimator is root-n consistent if $\lambda_n \to 0$. Further, it will be illustrated that such a root-n consistent estimator must satisfy $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and the proposed estimators perform as well as if $\boldsymbol{\beta}_{20} = \mathbf{0}$ were known, if $n^{1/2}\lambda_n \to \infty$. That is, the resulting estimator performs as well as the oracle estimator, which knows in advance $\boldsymbol{\beta}_{20} = \mathbf{0}$. In other words, the penalized likelihood estimator improves simultaneously the accuracy for estimating $(\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20})$.

In this chapter, we only consider the situations in which the parameter space for $\boldsymbol{\beta}$ is finite dimensional. It is assumed throughout this chapter that the penalty function $p_\lambda(\theta)$ has a second order continuous derivative at nonzero components of $\boldsymbol{\beta}_0$ and is nonnegative with $p_\lambda(0) = 0$. Denote

$$a_n = \max\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \tag{3.1.1}$$

$$b_n = \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \tag{3.1.2}$$

$$\Sigma = \operatorname{diag}\left\{p''_{\lambda_n}(|\beta_{10}|), \cdots, p''_{\lambda_n}(|\beta_{s0}|)\right\}, \tag{3.1.3}$$

$$\mathbf{b} = \left(p'_{\lambda_n}(|\beta_{10}|)\operatorname{sgn}(\beta_{10}), \cdots, p'_{\lambda_n}(|\beta_{s0}|)\operatorname{sgn}(\beta_{s0})\right)^T. \tag{3.1.4}$$

where s is the number of component of $\boldsymbol{\beta}_{10}$.

Asymptotic theory and oracle properties for penalized least-squares estimators and robust penalized least-squares estimators will be established in Sections 3.2 and 3.3, respectively. For fairly general likelihood settings, it will be shown in Section 3.4 that penalized likelihood estimators have oracle properties. The proposed penalized likelihood estimators in the context of generalized linear models will also be discussed in details in Section 3.4. The hard-thresholding penalty function and the SCAD may satisfy the conditions for root n consistency and oracle properties, therefore the resulting estimators are oracle estimators, while the L_1 penalty cannot satisfy the conditions for root n consistency or the oracle properties given in this chapter. It is conjectured that LASSO is not an oracle estimator. Some discussions on this aspect will be given in Section 3.5.

3.2 Penalized Least-squares Estimators

For the linear regression model

$$Y = \mathbf{x}^T\boldsymbol{\beta} + \varepsilon, \tag{3.2.1}$$

consistency and oracle property of penalized least-squares estimates will be established for the settings of both the random design and fixed design. Denote by $Q(\boldsymbol{\beta})$ the penalized least-squares function

$$\frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|) \equiv \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|).$$

First consider the case of random design.

Theorem 3.1 *Consider the model (3.2.1) with $E(\varepsilon|\mathbf{x}) = 0$, $var(\varepsilon|\mathbf{x}) = \sigma^2 < \infty$ and assume that the matrix $E\mathbf{x}\mathbf{x}^T$, $\mathbf{V}$ say, is finite and positive definite. Suppose that the observations $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ are independent and identically distributed. If $b_n \to 0$, then with probability tending to one, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| = O_P(n^{-1/2} + a_n)$, where a_n is given by (3.1.1)*

Proof: Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) > Q(\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon. \tag{3.2.2}$$

This implies with probability tending to one that there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n\mathbf{u} : \|\mathbf{u}\| \leq C\}$. Hence, there exists a local minimizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\alpha_n)$.

By the Strong Law of Large Numbers (SLLN), $\frac{1}{n}\mathbf{X}^T\mathbf{X} = \mathbf{V}\{1 + o_p(1)\}$. Using $p_{\lambda_n}(0) = 0$, it follows that

$$\begin{aligned} D_n(\mathbf{u}) &\equiv Q(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) - Q(\boldsymbol{\beta}_0) \\ &\geq \frac{1}{2}n\alpha_n^2\mathbf{u}^T\mathbf{V}\mathbf{u}\{1 + o_P(1)\} - \alpha_n\mathbf{u}^T\mathbf{X}^T(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta}_0) \\ &\quad + n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_j^0 + \alpha_n u_j|) - p_{\lambda_n}(|\beta_j^0|)\}. \end{aligned} \tag{3.2.3}$$

By the Central Limit Theorem (CLT) for summation of i.i.d random vectors, $n^{-1/2}\mathbf{X}^T(\mathbf{y} - \mathbf{X}^T\boldsymbol{\beta}_0) = O_P(1)$. Thus, the second term on the right hand side of (3.2.3) is of the order

$O_P(n^{1/2}\alpha_n) = O_P(n\alpha_n^2)$. Note that $\mathbf{V}$ is finite and positive definite. By choosing a sufficiently large C, the second term will be dominated by the first term, uniformly in $\|\mathbf{u}\| = C$. By Taylor's expansion, the third term on the right hand side of (3.2.3) becomes

$$\sum_{j=1}^{s} \left[n\alpha_n p'_{\lambda_n}(|\beta_j^0|)\mathrm{sgn}(\beta_j^0)u_j + n\alpha_n^2 p''_{\lambda_n}(|\beta_j^0|)u_j^2\{1+o(1)\} \right].$$

which is bounded by

$$\sqrt{s}n\alpha_n a_n\|\mathbf{u}\| + n\alpha_n^2 b_n\|\mathbf{u}\|^2.$$

This is also dominated by the first term of (3.2.3) as $b_n \to 0$. Hence, by choosing sufficiently large C, (3.2.2) holds. This completes the proof of the theorem.

It is clear from Theorem 1 that by choosing a proper λ_n, there exists a root-n consistent penalized likelihood estimator. We now show that this estimator must possess the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$, which is stated as follows.

Lemma 3.1 *Consider the model (3.2.1) with $E(\varepsilon|\mathbf{x}) = 0$, $var(\varepsilon|\mathbf{x}) = \sigma^2 < \infty$ and assume that the matrix $E\mathbf{x}\mathbf{x}^T$, $\mathbf{V}$ say, is finite and positive definite. Suppose that the observations $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ are independent and identically distributed. Assume that*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\theta) > 0. \tag{3.2.4}$$

If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to one, for any given $\boldsymbol{\beta}_1$ satisfying that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant C,

$$Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}\} = \max_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\}.$$

Proof: It suffices to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \cdots, d$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} > 0 \quad \text{for} \quad 0 < \beta_j < \varepsilon_n, \tag{3.2.5}$$

and

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for} \quad -\varepsilon_n < \beta_j < 0. \tag{3.2.6}$$

By some straightforward algebraic calculation, it follows that

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = -\mathbf{x}_{(j)}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0) + \mathbf{x}_{(j)}^T\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j),$$

where $\mathbf{x}_{(j)}$ is the j-th column of $\mathbf{X}$.

Note that by the CLT and the SLLN,

$$\frac{1}{n}\mathbf{x}_{(j)}^T(\mathbf{x} - \mathbf{X}\boldsymbol{\beta}_0) = O_P(n^{-1/2})$$

and

$$\frac{1}{n}\mathbf{x}_{(j)}^T\mathbf{X} = EX_j\mathbf{x} + o_P(1),$$

where X_j is the j-th component of $\mathbf{x}$.

By the assumption that $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n\{\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_P(n^{-1/2}/\lambda_n)\}.$$

Since $\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\lambda_n^{-1}p'_{\lambda_n}(\theta) > 0$ and $n^{-1/2}/\lambda_n \to 0$, the sign of the derivative is completely determined by that of β_j. Hence, (3.2.5) and (3.2.6) follow. This completes the proof.

Theorem 3.2 *(**Oracle Property**) Consider the model (3.2.1) with $E(\varepsilon|\mathbf{x}) = 0$, $var(\varepsilon|\mathbf{x}) = \sigma^2 < \infty$ and assume that the matrix $E\mathbf{x}\mathbf{x}^T$, $\mathbf{V}$ say, is finite and positive definite. Suppose that the observations $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ are independent and identically distributed. Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition (3.2.4). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the root n consistent local minimizer $\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2\end{pmatrix}$ in Theorem 3.1 must satisfy:*

*(i) (**Sparsity**) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;*

*(ii) (**Asymptotic normality**)*

$$\sqrt{n}(\mathbf{V}_{11} + \Sigma)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{V}_{11} + \Sigma)^{-1}\mathbf{b}\right\} \to N(\mathbf{0}, \sigma^2\mathbf{V}_{11})$$

in distribution, where $\mathbf{V}_{11}$ consists of the first s rows and columns of $\mathbf{V}$.

Proof: It follows by Lemma 3.1 that Part (i) holds. Now we prove Part (ii). It can be easily shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 3.1 that is a root n consistent local minimizer of $Q\{\begin{pmatrix}\boldsymbol{\beta}_1 \\ \mathbf{0}\end{pmatrix}\}$, regarded as a function of $\boldsymbol{\beta}_1$, and satisfying the following equations:

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1 \\ \mathbf{0}\end{pmatrix}} = 0, \quad \text{for} \quad j = 1, \cdots, s.$$

Note that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator,

$$\begin{aligned}\frac{\partial Q(\hat{\boldsymbol{\beta}})}{\partial \beta_j} &= -\mathbf{x}_{(j)}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0)+\mathbf{x}_{(j)}^T\mathbf{X}_{(1)}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})-np'_{\lambda_n}(|\hat{\beta}_j|)\\ &= -\mathbf{x}_{(j)}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0)+n\mathbf{V}_{(j)}^T(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})(1+o_P(1))\\ &\quad -n\left(p'_{\lambda_n}(|\beta_j^0|)\mathrm{sgn}(\beta_{j0})+\{p''_{\lambda_n}(|\beta_{j0}|)+o_P(1)\}(\hat{\beta}_j-\beta_{j0})\right).\end{aligned}$$

where $X_{(1)}$ consists of the first s columns of $\mathbf{X}$ and $\mathbf{V}_{(j)}$ is the jth-column of V_{11}. It follows by Slutsky's Theorem and the CLT that

$$\sqrt{n}(\mathbf{V}_{11}+\Sigma)\left\{\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}+(\mathbf{V}_{11}+\Sigma)^{-1}\mathbf{b}\right\}\to N\left(\mathbf{0},\sigma^2\mathbf{V}_{11}\right)$$

in distribution.

Now consider the settings in which the covariate vector $\mathbf{x}$ is from a fixed design. Then one may apply Hájek and Šidák Central Limit Theorem for the summation of weighted random vectors $\mathbf{x}_i^T(Y_i-\mathbf{x}_i^T\boldsymbol{\beta}_0)$ instead of the CLT for a summation of i.i.d random vectors. The following two conditions on the design matrix $\mathbf{X}$ are needed to guarantee the conditions of the Hájek and Šidák CLT,

$$\max_{1\le k\le n}\mathbf{x}_k^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_k\to 0 \quad \text{as} \quad n\to\infty \tag{3.2.7}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{X}^T\mathbf{X}=\mathbf{V}, \tag{3.2.8}$$

where $\mathbf{V}$ is finite and positive definite. See for example Sen and Singer (1993) for detailed arguments.

If the two conditions hold, we have the following two corollaries. Their proofs are almost the same as those of Theorems 3.1 and 3.2, and therefore are omitted.

Corollary 3.1 *Consider the model (3.2.1) and assume that the random errors are independent and identically distributed with zero mean and finite positive variance σ^2. If the conditions (3.2.7) and (3.2.8) hold and $b_n\to 0$, then with probability tending to one, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}||=O_P(n^{-1/2}+a_n)$, where a_n is given by (3.1.1).*

Corollary 3.2 *Consider the model (3.2.1) and assume that the random errors are independent and identically distributed with zero mean and finite positive variance σ^2. Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies the condition (3.2.4). If $\lambda_n\to 0$ and $\sqrt{n}\lambda_n\to\infty$ as $n\to\infty$,*

then with probability tending to 1, the root n consistent local minimizer $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ *in Corollary 3.1 must satisfy:*

*(i) (***Sparsity***)* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$*;*

*(ii) (***Asymptotic normality***)*

$$\sqrt{n}(\mathbf{V}_{11} + \Sigma)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{V}_{11} + \Sigma)^{-1}\mathbf{b}\right\} \to N\left(\mathbf{0}, \sigma^2 \mathbf{V}_{11}\right)$$

in distribution, where $\mathbf{V}_{11}$ *consists of the first* s *rows and columns of* $\mathbf{V}$*.*

As a consequence, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$ is

$$\frac{\sigma^2}{n}(\mathbf{V}_{11} + \Sigma)^{-1}\mathbf{V}_{11}(\mathbf{V}_{11} + \Sigma)^{-1},$$

which approximately equals $\frac{\sigma^2}{n}V_{11}^{-1}$ when n is large as $b_n \to 0$, which implies that $\Sigma \to \mathbf{0}$.

Now we compare the accuracy of the penalized least-squares estimator $\hat{\boldsymbol{\beta}}_1$ and the usual least-squares estimator $\hat{\boldsymbol{\beta}}_{\text{LS}}$. It is known that $\text{cov}(\hat{\boldsymbol{\beta}}_{\text{LS}}) \approx \frac{\sigma^2}{n}\mathbf{V}^{-1}$ as n is large. Let $\mathbf{U} = \mathbf{V}^{-1}$, and partition $\mathbf{U}$ and $\mathbf{V}$ as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix},$$

where $\mathbf{V}_{11}$ and $\mathbf{U}_{11}$ are $s \times s$ matrices. It is known (see for example Muirhead, 1982) that

$$\begin{aligned} \mathbf{U}_{11} &= (\mathbf{V}_{11} - \mathbf{V}_{12}\mathbf{V}_{12}\mathbf{V}_{21})^{-1} \\ &= \mathbf{V}_{11}^{-1} + \mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22}(\mathbf{V}_{22} + \mathbf{V}_{22}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}\mathbf{V}_{12}\mathbf{V}_{22})^{-1}\mathbf{V}_{22}\mathbf{V}_{21}\mathbf{V}_{11}^{-1}. \end{aligned}$$

Thus $\mathbf{U}_{11} - \mathbf{V}^{-1} \geq 0$ and equality holds if and only if $\mathbf{V}_{12} = 0$. Hence compared with the least-squares estimate, the penalized least-squares estimate also improves the accuracy for estimating $\boldsymbol{\beta}_{10}$.

3.3 Robust Penalized Least-squares Estimators

For the linear regression model (3.2.1), one minimizes the robust penalized least-squares function

$$Q(\boldsymbol{\beta}) \equiv \sum_{i=1}^{n} \psi(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|)$$

to obtain the robust penalized least-squares estimate. To established the consistency and the oracle property of the robust penalized the least-squares estimate, some regularity conditions on ψ-function are needed.

Condition on the ψ-function:

(B1) The function $\psi(x)$ has 2nd order continuous derivatives almost everywhere and satisfies

$$E\psi'(\varepsilon) = 0, \tag{3.3.1}$$

$$0 < \sigma^2 \equiv E\{\psi'(\varepsilon)\}^2 < \infty, \tag{3.3.2}$$

$$0 < \gamma \equiv E\{\psi''(\varepsilon)\} < \infty, \tag{3.3.3}$$

where ε is the random error in model (3.2.1).

Theorem 3.3 *For the model (3.2.1), assume that the random error ε and the covariate vector $\mathbf{x}$ are independent, and that the matrix $E\mathbf{x}\mathbf{x}^T$, $\mathbf{V}$ say, is finite and positive definite. Suppose that the observations $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ are independent and identically distributed. If $b_n \to 0$ and Condition (B1) holds, then with probability tending to one, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_P(n^{-1/2} + a_n)$ where a_n is given by (3.1.1).*

Proof: Note that ε and $\mathbf{x}$ are independent, it follows by the SLLN that

$$\frac{1}{n}\sum_{i=1}^{n} \psi''(\varepsilon_i)\mathbf{x}_i\mathbf{x}_i^T = \gamma\mathbf{V}\{1 + o_P(1)\}$$

Thus by Taylor's expansion, it holds that

$$\begin{aligned} & \sum_{i=1}^{n}[\psi\{Y_i - \mathbf{x}_i^T(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u})\} - \psi(Y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0)] \\ = \;& n\alpha_n^2\gamma\mathbf{u}^T\mathbf{V}\mathbf{u}\{1 + o_P(1)\} - \sqrt{n}\alpha_n\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi'(\varepsilon_i)\mathbf{x}_i^T\mathbf{u}. \end{aligned}$$

The proof is completed by following arguments similar to the proof of Theorem 3.1.

Lemma 3.2 *For the model (3.2.1), assume that the random error ε and covariate vector $\mathbf{x}$ are independent, and that the matrix $E\mathbf{x}\mathbf{x}^T$, $\mathbf{V}$ say, is finite and positive definite. Suppose that the observations $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ are independent and identically distributed. Suppose that the condition (3.2.4) and Condition (B1) hold. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then*

with probability tending to one, for any given $\boldsymbol{\beta}_1$ *satisfying that* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ *and for any constant* C,

$$Q\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\} = \max_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\}.$$

Proof: The proof is similar to that of Lemma 3.1. It suffices to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \cdots, d$, (3.2.5) and (3.2.6) hold. By some straightforward algebraic computation, it follows that

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = -\sum_{i=1}^{n} \psi'(\varepsilon_i) X_{ij} + \sum_{i=1}^{n} \psi''(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) X_{ij} \mathbf{x}^T (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + n p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j),$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$, and X_{ij} is the j-th component of $\mathbf{x}_i$. By the assumption that $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$ and the continuity of $\psi''(x)$, it follows that

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n \{\lambda_n^{-1} p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_P(n^{-1/2}/\lambda_n)\}.$$

Since $\liminf_{n\to\infty} \liminf_{\theta \to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\theta) > 0$ and $n^{-1/2}/\lambda_n \to 0$, the sign of the derivative is completely determined by that of β_j. Hence, (3.2.5) and (3.2.6) follow. This completes the proof.

Theorem 3.4 *For the model (3.2.1), assume that the random error* ε *and covariate vector* $\mathbf{x}$ *are independent, and that the matrix* $E\mathbf{x}\mathbf{x}^T$, $\mathbf{V}$ *say, is finite and positive definite. Suppose that the observations* $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ *are independent and identically distributed. Suppose that the condition (3.2.4) and the condition (B1) hold. If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to 1, the root n consistent local minimizer* $\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{pmatrix}$ *in Theorem 3.3 must satisfy:*

*(i) (***Sparsity***)* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

*(ii) (***Asymptotic normality***)*

$$\sqrt{n}(\gamma \mathbf{V}_{11} + \Sigma)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\gamma \mathbf{V}_{11} + \Sigma)^{-1}\mathbf{b}\right\} \to N(\mathbf{0}, \sigma^2 \mathbf{V}_{11})$$

in distribution, where $\mathbf{V}_{11}$ *consists of the first s rows and columns of* $\mathbf{V}$.

Proof: It follows by Lemma 3.2 that Part (i) holds. Now we prove Part (ii). It can be easily shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 3.3 that is a root n consistent local minimizer of

$Q\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\}$, regarded as a function of $\boldsymbol{\beta}_1$, and satisfies the following equations:

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}} = 0, \quad \text{for} \quad j = 1, \cdots, s.$$

Note that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator,

$$\begin{aligned}
\frac{\partial Q(\hat{\boldsymbol{\beta}})}{\partial \beta_j} &= -\sum_{i=1}^{n} \psi'(\varepsilon_i) X_{ij} + \sum_{i=1}^{n} \psi''(Y_i - \mathbf{x}_i^T \boldsymbol{\beta}^*) X_{ij} \mathbf{x}_{1,i}^T (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) + n p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) \\
= & -\sum_{i=1}^{n} \psi'(\varepsilon_i) X_{ij} + n\gamma \mathbf{V}_{(j)} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\{1 + o_P(1)\} \\
& + n\left(p'_{\lambda_n}(|\beta_j^0|)\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_P(1)\}(\hat{\beta}_j - \beta_{j0})\right),
\end{aligned}$$

where and $\mathbf{x}_{1,i}$ consists of the first s components of $\mathbf{x}_i$ and $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. It follows by Slutsky's Theorem and the CLT that

$$\sqrt{n}(\gamma \mathbf{V}_{11} + \Sigma)\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\gamma \mathbf{V}_{11} + \Sigma)^{-1}\mathbf{b}\right\} \to N\left(\mathbf{0}, \sigma^2 \mathbf{V}_{11}\right)$$

in distribution.

For the fixed design settings, the following corollaries hold. Their proofs are similar to those of Theorems 3.3 and 3.4, and therefore are omitted.

Corollary 3.3 *For the model (3.2.1), assume that the random errors are independent and identically distributed. If the condition (B1) and the conditions (3.2.7) and (3.2.8) hold, and $b_n \to 0$, then with probability tending to one, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}|| = O_P(n^{-1/2} + a_n)$, where a_n is given by (3.1.1).*

Corollary 3.4 *For the model (3.2.1), assume that the random errors are independent and identically distributed. Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies the condition (3.2.4), and Condition (B1) and the conditions (3.2.7) and (3.2.8) holds. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the root n consistent local minimizer $\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\beta}}_2\end{pmatrix}$ in Corollary 3.3 must satisfy:*

*(i) (***Sparsity***) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;*

(ii) (**Asymptotic normality**)

$$\sqrt{n}(\gamma \mathbf{V}_{11} + \Sigma) \left\{ \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\gamma \mathbf{V}_{11} + \Sigma)^{-1} \mathbf{b} \right\} \to N\left(\mathbf{0}, \sigma^2 \mathbf{V}_{11}\right)$$

in distribution, where $\mathbf{V}_{11}$ *consists of the first* s *rows and columns of* $\mathbf{V}$.

3.4 Penalized Likelihood Estimators

It this section, first consistency and oracle properties will be established in fairly general likelihood settings, and then penalized likelihood estimators will be discussed in the context of generalized linear models in detail. To facilitate the presentation, it is assumed that penalization is applied to every component of $\boldsymbol{\beta}$. However, there is no extra difficulty to extend it to the case where some components (e.g. variance in the linear models) are not penalized.

Set $\mathbf{v}_i = (\mathbf{x}_i, Y_i)$, $i = 1, \cdots, n$ and denote by Ω the parameter space for $\boldsymbol{\beta}$. We first state some regularity conditions.

Regularity Conditions:

(C1) The observations $\mathbf{v}_i$ are independent and identically distributed with probability density $f(\mathbf{v}, \boldsymbol{\beta})$ with respect to some measure μ. $f(\mathbf{v}, \boldsymbol{\beta})$ has a common support and the model is identifiable. Furthermore, the first and second logarithmic derivatives of f satisfy the equations

$$E_{\boldsymbol{\beta}} \left[\frac{\partial \log f(\mathbf{v}, \boldsymbol{\beta})}{\partial \beta_j} \right] = 0 \quad \text{for} \quad j = 1, \cdots, d$$

and

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

(C2) The Fisher information matrix

$$I(\boldsymbol{\beta}) = E \left\{ \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{v}, \boldsymbol{\beta}) \right] \left[\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{v}, \boldsymbol{\beta}) \right]^T \right\}$$

is finite and positive definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

(C3) There exists an open subset ω of Ω containing the true parameter point $\boldsymbol{\beta}_0$ such that for almost all $\mathbf{v}$ the density $f(\mathbf{v}, \boldsymbol{\beta})$ admits all third derivatives $\frac{\partial f(\mathbf{v}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_k \partial \beta_l}$ for all $\boldsymbol{\beta} \in \Omega$. Further there exist functions M_{jkl} such that

$$\left| \frac{\partial^3}{\partial \beta_j \partial \beta_k \partial \beta_l} \log f(\mathbf{v}, \boldsymbol{\beta}) \right| \leq M_{jkl}(\mathbf{v}) \quad \text{for all} \quad \boldsymbol{\beta} \in \Omega,$$

where $m_{jkl} = E_{\boldsymbol{\beta}_0}[M_{jkl}(\mathbf{v})] < \infty$ for j, k, l.

These regularity conditions guarantee asymptotic normality of the ordinary maximum likelihood estimates. See for example Lehmann (1983). When the density $f(\mathbf{v}, \boldsymbol{\beta})$ does not admit all third derivatives, Condition (C3) can be relaxed to

(C3')

$$E_{\boldsymbol{\beta}_0} \left\{ \sup_{\|\mathbf{h}\| < \delta} \| \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log f(\mathbf{v}, \boldsymbol{\beta}_0 + \mathbf{h}) - \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log f(\mathbf{v}, \boldsymbol{\beta}_0) \| \right\} \to 0$$

as $\delta \to 0$.

See for example Sen and Singer (1993) for details.

Let $L(\boldsymbol{\beta})$ be the log-likelihood function of observations $\mathbf{v}_1, \cdots, \mathbf{v}_n$ and let $Q(\boldsymbol{\beta})$ be the penalized likelihood function $L(\boldsymbol{\beta}) - n \sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|)$.

Theorem 3.5 *Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be independent and identically distributed each with a density $f(\mathbf{v}, \boldsymbol{\beta})$ (with respect to a measure μ) which satisfies Conditions (C1)–(C3). If $b_n \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$, where a_n is given by (3.1.1).*

Proof: Let $\alpha_n = n^{-1/2} + a_n$. We want to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P \left\{ \sup_{\|\mathbf{u}\| = C} Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) < Q(\boldsymbol{\beta}_0) \right\} \geq 1 - \varepsilon. \tag{3.4.1}$$

This implies with probability at least $1 - \varepsilon$ that there exists a local maximum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\alpha_n)$.

Using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D_n(\mathbf{u}) &\equiv Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\beta}_0) \\ &\leq L(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - L(\boldsymbol{\beta}_0) - n \sum_{j=1}^{s} \{p_{\lambda_n}(|\beta_j^0 + \alpha_n u_j|) - p_{\lambda_n}(|\beta_j^0|)\}, \end{aligned}$$

Let $L'(\boldsymbol{\beta}_0)$ be the gradient vector of L. By the standard argument based on the Taylor expansion of the likelihood function, it follows that

$$\begin{aligned} D_n(\mathbf{u}) &\leq \alpha_n L'(\boldsymbol{\beta}_0)^T \mathbf{u} - \frac{1}{2} \mathbf{u}^T I(\boldsymbol{\beta}_0) \mathbf{u} n \alpha_n^2 \{1 + o_P(1)\} \\ &\quad - \sum_{j=1}^{s} \left[n \alpha_n p'_{\lambda_n}(|\beta_j^0|) \mathrm{sgn}(\beta_j^0) u_j + n \alpha_n^2 p''_{\lambda_n}(|\beta_j^0|) u_j^2 \{1 + o(1)\} \right]. \end{aligned} \tag{3.4.2}$$

Note that $n^{-1/2} L'(\boldsymbol{\beta}_0) = O_P(1)$. Thus, the first term on the right hand side of (3.4.2) is of the order $O_P(n^{1/2}\alpha_n) = O_P(n\alpha_n^2)$. By a choosing sufficient large C, the second term will dominate the first term, uniformly in $\|\mathbf{u}\| = C$. Note that the third term in (3.4.2) is bounded by

$$\sqrt{s} n \alpha_n a_n \|\mathbf{u}\| + n \alpha_n^2 b_n \mathbf{u}\|^2.$$

This is also dominated by the second term of (3.4.2). Hence, by choosing a sufficiently large C, (3.4.1) holds. This completes the proof of the theorem.

It is clear from Theorem 3.5 that by choosing a proper λ_n, there exists a root-n consistent penalized likelihood estimator. We now show that this estimator must possess the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$, which is stated as follows.

Lemma 3.3 *Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be independent and identically distributed each with a density $f(\mathbf{v}, \boldsymbol{\beta})$ which satisfies Conditions (C1)–(C3). Assume that the penalty function satisfies condition (3.2.4). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant C,*

$$Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}\} = \max_{\|\boldsymbol{\beta}_2\| \leq C n^{-1/2}} Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\}.$$

Proof: It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$ and for some small $\varepsilon_n = C n^{-1/2}$ and $j = s+1, \cdots, d$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for} \quad 0 < \beta_j < \varepsilon_n, \tag{3.4.3}$$

and

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} > 0 \quad \text{for} \quad -\varepsilon_n < \beta_j < 0. \tag{3.4.4}$$

To show (3.4.3), by Taylor's expansion, we have

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} - n p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) \\
&= \frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{l0}) \\
&\quad + \sum_{l=1}^{d}\sum_{k=1}^{d} \frac{\partial^3 L(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_l \beta_k}(\beta_l - \beta_{l0})(\beta_k - \beta_{k0}) - n p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j),
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that by standard arguments

$$n^{-1}\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} = O_P(n^{-1/2})$$

and

$$\frac{1}{n}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} = E\left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}\right\} + o_P(1).$$

By the assumption that $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$, we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n\{-\lambda_n^{-1} p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) + O_P(n^{-1/2}/\lambda_n)\}.$$

Since $\liminf_{n\to\infty} \liminf_{\theta\to 0^+} \lambda_n^{-1} p'_{\lambda_n}(\theta) > 0$ and $n^{-1/2}/\lambda_n \to 0$, the sign of the derivative is completely determined by that of β_j. Hence, (3.4.3) and (3.4.4) follow. This completes the proof.

Theorem 3.6 *(**Oracle property**) Let $\mathbf{v}_1, \cdots, \mathbf{v}_n$ be independent and identically distributed each with a density $f(\mathbf{v}, \boldsymbol{\beta})$ satisfying Conditions (C1)–(C3). Assume that the penalty function $p_{\lambda_n}(|\theta|)$ satisfies condition (3.2.4). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the root n consistent local maximizers $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ in Theorem 3.5 must satisfy:*

*(i) (**Sparsity**) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;*

*(ii) (**Asymptotic normality**)*

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

in distribution, where $I_1(\boldsymbol{\beta}_{10}) = I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$, the Fisher information knowing $\boldsymbol{\beta}_2 = \mathbf{0}$.

Proof: It follows by Lemma 1 that Part (i) holds. Now we prove Part (ii). It can be easily shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 1 that is a root n consistent local maximizer of $Q\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\}$, regarded as a function of $\boldsymbol{\beta}_1$, and satisfies the likelihood equations:

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}} = 0, \quad \text{for} \quad j = 1, \cdots, s. \tag{3.4.5}$$

Note that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator,

$$\begin{aligned}
&\left.\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}} - n p'_{\lambda_n}(|\hat{\beta}_j|) \\
= \;& \frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} + o_P(1)\right\}(\hat{\beta}_l - \beta_{l0}) \\
& -n\left(p'_{\lambda_n}(|\beta_j^0|)\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_P(1)\}(\hat{\beta}_j - \beta_{j0})\right).
\end{aligned}$$

It follows by Slutsky's Theorem and the CLT that

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

in distribution.

As a consequence, the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_1$ is

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

which approximately equals $\frac{1}{n}I_1^{-1}(\boldsymbol{\beta}_{10})$ for the thresholding penalties discussed in Section 2 if λ_n tends to 0.

It has been shown that there exists a penalized likelihood estimator that converges at rate $O_P(n^{-1/2} + a_n)$. This implies that for the hard thresholding and SCAD penalty functions, the penalized likelihood estimator is root-n consistent if $\lambda_n \to 0$. It also has been demonstrated that such a root-n consistent estimator must satisfy $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and $\hat{\boldsymbol{\beta}}_1$ is asymptotically normal with covariance matrix $I_1^{-1}(\boldsymbol{\beta}_{10})$, if $n^{1/2}\lambda_n \to \infty$. This implies that the penalized likelihood estimator performs as well as if $\boldsymbol{\beta}_{20} = \mathbf{0}$ were known. In the similar language of Donoho and Johnstone (1994a), the resulting estimator performs as well as the oracle estimator, which knows in advance $\boldsymbol{\beta}_{20} = \mathbf{0}$. In other words, the penalized likelihood estimator improves simultaneously the accuracy for estimating $(\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20})$.

Now we apply the results in Theorems 3.5 and 3.6 for penalized likelihood estimators in the context of generalized linear models. Generalized linear models, due to Nelder and Wedderburn (1972), have been systematically studied by McMullagh and Nelder (1989). In the context of generalized linear models, the conditional distribution of the response variable Y given $\mathbf{x}$ belongs to the exponential family with density function

$$f(y, \theta(\mathbf{x}), \phi) = \exp\left([\theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}]/a(\phi) + c(y, \phi)\right). \tag{3.4.6}$$

In parametric generalized linear models, it is usual to model a transform of the regression function $E(Y|\mathbf{x})$ as linear, that is

$$g\{E(Y|\mathbf{x})\} = \mathbf{x}^T\boldsymbol{\beta},$$

where g is a known link function. If $g = (b')^{-1}$, then g is called the canonical link. It is well-known that the mean function of Y is

$$\mu(\mathbf{x}^T\boldsymbol{\beta}) = E(Y|\mathbf{x}) = b'\{\theta(\mathbf{x})\},$$

and the variance function is

$$v(\mathbf{x}^T\boldsymbol{\beta}) = \mathrm{var}(Y|\mathbf{x})/a(\phi) = b''\{\theta(\mathbf{x})\}.$$

The log-likelihood function of $(\mathbf{x}, Y)$ is

$$\log\{f(y, \theta(\mathbf{x}), \phi)\} = y(g \circ \mu)^{-1}(\mathbf{x}^T\boldsymbol{\beta}) - b\{(g \circ \mu)^{-1}(\mathbf{x}^T\boldsymbol{\beta})\} + C(\mathbf{x}, y),$$

where $C(\mathbf{x}, y)$ does not depend on unknown parameters.

Thus the Hessian matrix of the log-likelihood function is

$$\frac{\partial^2 \log f}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = w_1(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T + \{y - \mu(\mathbf{x}^T\boldsymbol{\beta})\}w_2(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T,$$

where

$$w_1(\mathbf{x}^T\boldsymbol{\beta}) = \frac{[g'\{\mu(\mathbf{x}^T\boldsymbol{\beta}\}]^2}{v(\mathbf{x}^T\boldsymbol{\beta})},$$

and

$$w_2(\mathbf{x}^T\boldsymbol{\beta}) = \frac{g''\{\mu(\mathbf{x}^T\boldsymbol{\beta})\}}{[g'\{\mu(\mathbf{x}^T\boldsymbol{\beta})\}]^2} + \frac{b'''\{(g \circ \mu)^{-1}(\mathbf{x}^T\boldsymbol{\beta})\}}{[g'\{\mu(\mathbf{x}^T\boldsymbol{\beta})\}]^2\{v(\mathbf{x}^T\boldsymbol{\beta})\}^3}.$$

Thus the Fisher information matrix is

$$I(\boldsymbol{\beta}) = -E\left\{w_1(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T\right\}. \tag{3.4.7}$$

Define

$$w_3(\mathbf{x}^T\boldsymbol{\beta}) = \mu(\mathbf{x}^T\boldsymbol{\beta})w_2(\mathbf{x}^T\boldsymbol{\beta})$$

The conditions (3.4.8) and (3.4.9) below are sufficient for Condition (C3'):

$$E_{\boldsymbol{\beta}_0}\left\{\sup_{\mathbf{h}:\|\mathbf{h}\|\le\delta}|w_k\{\mathbf{x}^T(\boldsymbol{\beta}_0+\mathbf{h})\} - w_k(\mathbf{x}^T\boldsymbol{\beta}_0)|\|\mathbf{x}\|^2\right\} \to 0 \tag{3.4.8}$$

as $\delta \to 0$ for $k = 1, 3$; and

$$E_{\boldsymbol{\beta}_0}\left\{\sup_{\mathbf{h}:\|\mathbf{h}\|\le\delta}|Y||w_2\{\mathbf{x}^T(\boldsymbol{\beta}_0+\mathbf{h})\} - w_2(\mathbf{x}^T\boldsymbol{\beta}_0)|\|\mathbf{x}\|^2\right\} \to 0 \tag{3.4.9}$$

as $\delta \to 0$.

For generalized linear models with canonical link, the Hessian matrix of the log-likelihood function is

$$\frac{\partial^2 \log f}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = -b''(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T/a(\phi).$$

So the Fisher information matrix is

$$I(\boldsymbol{\beta}) = Eb''(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T/a(\phi).$$

The following condition is sufficient for Condition (C3') when the link function g is the canonical link.

$$E_{\boldsymbol{\beta}_0}\left\{\sup_{\mathbf{h}:\|\mathbf{h}\|\le\delta}|b''\{\mathbf{x}^T(\boldsymbol{\beta}_0+\mathbf{h})\} - b''(\mathbf{x}^T\boldsymbol{\beta}_0)|\|\mathbf{x}\|^2\right\} \to 0 \tag{3.4.10}$$

as $\delta \to 0$.

It follows directly from Theorems 3.5 and 3.6 that the following corollaries hold.

Corollary 3.5 *Let $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ be independent and identically distributed each with a density (3.4.6) with finite and positive definite Fisher information matrix $I(\boldsymbol{\beta}_0)$ in (3.4.7). If the conditions (3.4.8) and (3.4.9) holds, and $b_n \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$, where a_n is given by (3.1.1).*

Corollary 3.6 *Let $(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)$ be independent and identically distributed each with a density (3.4.6) with finite and positive definite Fisher information matrix $I(\boldsymbol{\beta}_0)$ in (3.4.7). Suppose that If the conditions (3.4.8) and (3.4.9) holds, and the penalty function satisfies the condition (3.2.4). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the root n consistent local maximizer $\hat{\boldsymbol{\beta}} = \begin{pmatrix}\hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2\end{pmatrix}$ in Corollary 3.5 must satisfy:*

(i) (**Sparsity**) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

(ii) (**Asymptotic normality**)

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

in distribution, where $I_1(\boldsymbol{\beta}_{10}) = I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$, *the Fisher information knowing* $\boldsymbol{\beta}_2 = \mathbf{0}$.

For the settings in which the covariate $\mathbf{x}$ is the fixed design, the following conditions guarantee the results in Corollaries 3.5 and 3.6 hold.

Regularity conditions:

(D1)

$$\lim_{n\to\infty} \frac{1}{n}\left(\frac{[g'\{\mu(\mathbf{x}_i^T\boldsymbol{\beta}_0\}]^2}{v(\mathbf{x}_i^T\boldsymbol{\beta}_0)}\mathbf{x}_i\mathbf{x}_i^T\right) = I(\boldsymbol{\beta}_0)$$

which is finite and positive definite.

(D2)

$$\lim_{n\to\infty} n^{-2}\sum_{i=1}^{n} v(\mathbf{x}_i^T\boldsymbol{\beta}_0)w_2(\mathbf{x}_i^T\boldsymbol{\beta}_0)\|\mathbf{x}_i\|^4 \to 0.$$

(D3) For $k = 1, 3$,

$$\sup_{\|\mathbf{h}\|<\delta}\left(\frac{1}{n}\sum_{i=1}^{n}|w_k\{\mathbf{x}_i^T(\boldsymbol{\beta}_0 + \mathbf{h})\} - w_k(\mathbf{x}_i^T\boldsymbol{\beta}_0)|\|\mathbf{x}_i\|^2\right) \to 0$$

as $\delta \to 0$, and

$$E_{\boldsymbol{\beta}_0}\sup_{\|\mathbf{h}\|<\delta}\left(\frac{1}{n}\sum_{i=1}^{n}|Y_i||w_2\{\mathbf{x}_i^T(\boldsymbol{\beta}_0 + \mathbf{h})\} - w_2(\mathbf{x}_i^T\boldsymbol{\beta}_0)|\|\mathbf{x}_i\|^2\right) \to 0$$

as $\delta \to 0$.

Condition (D2) implies that

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - \mu(\mathbf{x}_i\boldsymbol{\beta}_0)\}w_2(\mathbf{x}_i^T\boldsymbol{\beta}_0)\mathbf{x}_i\mathbf{x}_i^T \to 0$$

in probability by Chebyshev inequality. Therefore Condition (D1) implies the Fisher information $I(\boldsymbol{\beta}_0)$ is finite and positive definite. Condition (D3) enables us to show that Condition (C3') holds. As a consequence of Theorems 3.5 and 3.6, the following corollaries hold.

Corollary 3.7 *Let $Y_1, \cdots, Y_n$ be independent observations from the population with a density (3.4.6). Suppose that Conditions (D1)–(D3) hold. If $b_n \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$, where a_n is given by (3.1.1).*

Corollary 3.8 *Let $Y_1, \cdots, Y_n$ be independent observations from the population with a density (3.4.6). Suppose that Conditions (D1)–(D3) hold, and the penalty function satisfies the condition (3.2.4). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to 1, the root n consistent local maximizer $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ in Corollary 3.7 must satisfy:*

*(i) (***Sparsity***)* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$*;*

*(ii) (***Asymptotic normality***)*

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

in distribution, where $I_1(\boldsymbol{\beta}_{10}) = I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$, the Fisher information knowing $\boldsymbol{\beta}_2 = \mathbf{0}$.

3.5 Discussion and Conclusion

For the hard, mixture and SCAD thresholding penalty functions, if $\lambda_n \to 0$, then $a_n = 0$. Hence, by Theorem 3.6, when $\sqrt{n}\lambda_n \to \infty$, their corresponding penalized likelihood estimators possess the oracle property and perform as well as the maximum likelihood estimates for estimating $\boldsymbol{\beta}_1$ knowing $\boldsymbol{\beta}_2 = \mathbf{0}$. They are more efficient than the maximum likelihood estimator for estimating $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. However, for the L_1 penalty, $a_n = \lambda_n$. Hence, root-n consistency requires that $\lambda_n = O_P(n^{-1/2})$. On the other hand, the oracle property requires that $\sqrt{n}\lambda_n \to \infty$. These two conditions for LASSO cannot be satisfied simultaneously. Indeed, for the L_1 penalty, we conjecture that the oracle property does not hold. But for the L_p-penalty with $p < 1$, the oracle property continues to hold with suitable choice of λ_n.

This chapter demonstrates how the rates of convergence for the penalized likelihood estimators depend on the regularization parameter. It has been shown that the penalized likelihood estimators perform as well as the oracle procedure in terms of selecting the correct model, when the regularization parameter is appropriately chosen. In other words, when the true parameters have some zero components, they are estimated as 0 with probability tending to one, and the non-zero components are estimated as well as if the correct submodel were known.

This improves the accuracy for estimating not only the null components, but also the non-null components. In short, the penalized likelihood estimators work as well as if the correct submodel were known in advance. The significance of this is that the proposed procedures outperform the maximum likelihood estimator and perform as well as we hope. This is very analogous to the super-efficiency phenomenon, for instance, Hodges example in page 405 of Lehmann (1983).

Chapter 4

Variable Selection in Survival Data Analysis

4.1 Introduction

The problem of analyzing time to event data arises in a number of applied fields, such as medicine, engineering, biology, epidemiology, economics, public health and demography. A common feature of time to event data is that it contains either censored or truncated observations. Censored data arise when a subject's life length is known to occur only in a certain period of time. Possible censoring schemes are *right censoring*, where all some subjects are still alive at a given time, *left censoring* when some individuals have experienced the event of interest prior to the start of the study, or *interval censoring*, where the only information is that the event occurs within some interval. Truncation schemes are *left truncation*, where only individuals who survival a sufficient time are included in the sample and *right truncation*, where only individuals who have experienced the event by a specified time are included in the sample. Detailed definitions of censoring and truncation scheme can be found for example in Chapter 3 of Klein and Moeschberger (1997). Only the right censoring scheme is considered in this chapter.

In the absence of any covariates, one may use the product-limit estimator (see Kaplan and Meier, 1958) to estimate the survival function, and use the Nelson-Aalen estimator (Nelson, 1972 and Aalen 1978) to estimate the cumulative hazard function. Nonparametric regression techniques, such as kernel smoothing, can be applied for estimating the hazard function, see

for example Ramlau-Hansen (1983), Fan and Gijbels (1994) and González-Manteiga, Cao and Marron (1996). Bayesian nonparametric methods can also be used to estimate the hazard function.

In most studies, researchers are interested in comparing hazard rates among different groups, such as age and gender. It is usual to model the hazard function using some parametric models, for example, the accelerated failure-time model and the proportional hazards model (see Klein and Moeschberger, 1997 for details). It is often assumed that hazard functions depend on covariates through linear combination of covariates, $\mathbf{x}^T\boldsymbol{\beta}$ say, where $\mathbf{x}$ is the covariate vector. In many situations, there are a number of covariates available to be potential predictors. Consider the situation where all explanatory variables are on an equal footing, and our target is to select a subset of significant variables upon which the hazard function depends. Best subset selection and stepwise deletion are used to select significant variables in this context. Motivated by Lindley (1968), Faraggi and Simon (1998) proposed a Bayesian variable selection method for censored survival data. However the methods aforementioned are discrete processes in the sense that a predictor variable is either retained or deleted. Tibshirani (1997) applied the idea of LASSO for Cox's proportional hazards model. LASSO, however, will cause large model bias if the thresholding parameter is large. The proposed penalized likelihood approaches in Chapter 2 can be directly applied for parametric models in the context of survival data analysis. The purpose of this chapter is to extend the ideas of variable selection via nonconcave penalized likelihood to semi-parametric models. We are particularly interested in Cox's proportional hazards model and the frailty model, two commonly used semi-parametric models in the context of survival data analysis.

It will be shown in Section 4.2 that the penalized likelihood of Cox's regression model is equivalent to penalized partial likelihood. The rates of convergence and oracle properties will be established for the proposed penalized partial likelihood estimators in Section 4.2.2. Furthermore variable selection for Cox's proportional hazards frailty model will be studied in Section 4.3. There are a number of papers concerned with the estimation problems of regression coefficients in frailty models in the recent literature. Lee *et al* (1992) suggested the use of pseudo-likelihood approach to estimate the regression coefficients when the number in each group is small. Nielsen *et al* (1992) applied the EM algorithm to the gamma-frailty model (see Section 4.3 below for definition). Sinha (1998) proposed a posterior likelihood method for the gamma frailty model. It seems that there is no satisfactory method for estimating regression coefficients in semi-parametric frailty models. Therefore a new approach to estimating regres-

sion coefficients in Cox's proportional hazard frailty models is proposed, motivated by the idea of backfitting in the context of nonparametric regression. The proposed approach can be used to derive a unified variable selection approach via nonconcave penalized likelihood for the semi-parametric frailty model. Numerical simulations are conducted to compare the performance of the proposed approach and the pseudo-likelihood estimator in Section 4.4. Comparisons among the performance of the proposed variable selection procedures and best subset variable selection are also given in Section 4.4.

4.2 Proportional Hazards Models

Notation in this chapter is consistent with that in Sections 5.2 and 5.3 of Fan and Gijbels (1996). Let T, C and $\mathbf{x}$ be respectively the survival time, the censoring time and their associated covariates. Correspondingly, let $Z = \min\{T, C\}$ be the observed time and $\delta = I(T \leq C)$ be the censoring indicator. It will be assumed that T and C are conditionally independent given $\mathbf{x}$ and that the censoring mechanism is noninformative. When the observed data $\{(\mathbf{x}_i, Z_i, \delta_i) : i = 1, \cdots, n\}$ are an independent and identically distributed random sample from a certain population $(\mathbf{x}, Z, \delta)$, a complete likelihood of the data is given by

$$L = \prod_u f(Z_i|\mathbf{x}_i) \prod_c \overline{F}(Z_i|\mathbf{x}_i) = \prod_u h(Z_i|\mathbf{x}_i) \prod_{i=1}^n \overline{F}(Z_i|\mathbf{x}_i), \tag{4.2.1}$$

where the subscripts c and u denote the product of the censored and uncensored data respectively, and $f(t|\mathbf{x})$, $\overline{F}(t|\mathbf{x})$ and $h(t|\mathbf{x})$ are the conditional density function, conditional survival function of t and the corresponding hazard function at t given $\mathbf{x}$. Statistical inference in this chapter will be based on the likelihood function (4.2.1). To present the complete likelihood function of Cox's proportional hazards model, more notations are needed. Let $t_1^0 < \cdots < t_N^0$ denote the ordered observed failure times. Let (j) provide the label for the item failing at t_j^0 so that the covariates associated with the N failures are $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(N)}$. Let R_j denote the risk set at time t_{j-1}^0 :

$$R_j = \{i : Z_i \geq t_j^0\}.$$

Consider the proportional hazards model,

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}).$$

The likelihood in (4.2.1) becomes

$$L = \prod_{i=1}^{N} h_0(Z_{(i)}) \exp(\mathbf{x}_{(i)}^T \boldsymbol{\beta}) \prod_{i=1}^{n} \exp\{-H_0(Z_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\},$$

where $H_0(\cdot)$ is the cumulative baseline hazard function. If the baseline hazard function has a parametric functional form, $h_0(\boldsymbol{\theta}, \cdot)$ says. Then the corresponding penalized likelihood function is

$$\sum_{i=1}^{N} [\log\{h_0(\boldsymbol{\theta}, Z_{(i)})\} + \mathbf{x}_{(i)}^T \boldsymbol{\beta}] - \sum_{i=1}^{n} \{H_0(\boldsymbol{\theta}, Z_i) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\} - n \sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{4.2.2}$$

Maximizing (4.2.2) with respect to $(\boldsymbol{\theta}, \boldsymbol{\beta})$ yields a maximum penalized likelihood estimator.

4.2.1 Cox's proportional hazard model and penalized partial likelihood

In Cox's proportional hazards model, the baseline hazard function is unknown and has not been parameterized either. Following the idea of Breslow (1972), consider a "least informative" nonparametric model for $H_0(\cdot)$, i.e. $H_0(t)$ has a possible jump h_j at the observed failure time t_j^0. More precisely

$$H_0(t) = \sum_{j=1}^{N} h_j I(t_j^0 \leq t).$$

Then

$$H_0(Z_i) = \sum_{j=1}^{N} h_j I(i \in R_j). \tag{4.2.3}$$

Using (4.2.3), the log-likelihood function becomes

$$\sum_{j=1}^{N} \{\log(h_j) + \mathbf{x}_{(j)}^T \boldsymbol{\beta}\} - \sum_{i=1}^{n} \{\sum_{j=1}^{N} h_j I(i \in R_j) \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}. \tag{4.2.4}$$

Taking the derivative with respect to h_j and setting it zero gives

$$\hat{h}_j = \{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}^{-1}.$$

Substituting $\hat{h}_j$ into (4.2.4), the profile likelihood of the likelihood function (4.2.1) becomes a partial likelihood (Cox, 1975). Therefore the penalized likelihood for Cox's proportional hazards model becomes the penalized partial likelihood

$$\sum_{j=1}^{N} [\mathbf{x}_{(j)}^T \boldsymbol{\beta} - \log\{\sum_{i \in R_j} \exp(\mathbf{x}_i^T \boldsymbol{\beta})\}] - N - n \sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{4.2.5}$$

To obtain the penalized likelihood estimate of $\boldsymbol{\beta}$ for Cox's proportional hazards model, we maximize (4.2.5) with respect to $\boldsymbol{\beta}$.

4.2.2 Sampling properties and oracle properties

Strong consistency and asymptotic normality of the maximum partial likelihood estimator of the regression parameter in Cox's regression model have been established in 1980's. Tsiatis (1981) proved the strong consistency and asymptotic normality under the assumptions that the covariate is a random scalar variable and the censoring random variable is bounded. He also suggested how to extend his idea to the case of multidimensional covariates. Bailey (1983) showed the asymptotic normality of the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$ when the covariate $\mathbf{x}$ is the fixed design. Wong (1986) set up a general theory for partial likelihood estimates. Slud (1982) studied consistency and efficiency of maximum partial likelihood estimates for the two-sample survival problem.

In this section, the asymptotic theory will be established for the proposed penalized likelihood estimator. Let

$$\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T.$$

Without loss of generality, assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$. Denote by s the number of components of $\boldsymbol{\beta}_1$, and let

$$a_n = \max\{|p'_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \tag{4.2.6}$$

and

$$b_n = \max\{|p''_{\lambda_n}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}. \tag{4.2.7}$$

It will be shown that there exists a penalized partial likelihood estimator that converges at rate $O_P(n^{-1/2} + a_n)$. Oracle properties for penalized partial likelihood estimator will also be established. In this section, we only state theoretical results, their proofs will be given in Section 4.5.

The following theorem illustrates how the rates of convergence for the penalized partial likelihood estimators depend on the regularization parameter. Let $\ell(\boldsymbol{\beta})$ denote the log-partial likelihood function

$$\sum_{j=1}^{N}[\mathbf{x}_{(j)}^T\boldsymbol{\beta} - \log\{\sum_{i\in R_j}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}]$$

and let $Q(\boldsymbol{\beta})$ be the penalized partial likelihood function $\ell(\boldsymbol{\beta}) - \sum_{j=1}^{d} p_\lambda(|\beta_j|)$.

Theorem 4.1 *Assume that* $(\mathbf{x}_1, T_1, C_1), \cdots, (\mathbf{x}_n, T_n, C_n)$ *are independent and identically distributed according to the population* $(\mathbf{x}, T, C)$, T *and* C *are conditionally independent given* $\mathbf{x}$, *and* $E\{\exp(2\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T\}$ *is bounded uniformly in a neighborhood of* $\boldsymbol{\beta}_0$. *If* $b_n \to 0$, *then there*

exists a local maximizer $\hat{\boldsymbol{\beta}}$ *of* $Q(\boldsymbol{\beta})$ *such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$, *where* a_n *is given by (4.2.6).*

It is clear from Theorem 1 that by choosing a proper λ_n, there exists a root-n consistent penalized partial likelihood estimator. We now show that this estimator must possess the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$, which is stated as follows.

Lemma 4.1 *Assume that* $(\mathbf{x}_1, T_1, C_1), \cdots, (\mathbf{x}_n, T_n, C_n)$ *are independent and identically distributed according to the population* $(\mathbf{x}, T, C)$, T *and* C *are conditionally independent given* $\mathbf{x}$, *and* $E\{\exp(2\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T\}$ *is bounded uniformly in a neighborhood of* $\boldsymbol{\beta}_0$. *Assume that*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0. \tag{4.2.8}$$

If $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to 1, for any given* $\boldsymbol{\beta}_1$ *satisfying* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ *and any constant* C,

$$Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}\} = \max_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\}.$$

Denote

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

and

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

where s is the number of components of $\boldsymbol{\beta}_{10}$.

Theorem 4.2 *(***Oracle property***) Assume that* $(\mathbf{x}_1, T_1, C_1), \cdots, (\mathbf{x}_n, T_n, C_n)$ *are independent and identically distributed according to the population* $(\mathbf{x}, T, C)$, T *and* C *are conditionally independent given* $\mathbf{x}$, *and* $E\{\exp(2\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T\}$ *is bounded uniformly in a neighborhood of* $\boldsymbol{\beta}_0$. *Assume that the penalty function* $p_{\lambda_n}(|\theta|)$ *satisfies the condition (4.2.8). If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$ *as* $n \to \infty$, *then with probability tending to 1, the root n consistent local maximizers* $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}$ *in Theorem 4.1 must satisfy:*

*(i) (***Sparsity***)* $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$;

(ii) (**Asymptotic normality**)

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

in distribution, where $I_1(\boldsymbol{\beta}_{10}) = I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$, *the Fisher information matrix of partial likelihood knowing* $\boldsymbol{\beta}_2 = \mathbf{0}$. *That is*

$$I_1(\boldsymbol{\beta}_{10}) = -E\left\{\frac{\partial^2 \ell((\boldsymbol{\beta}_{10}, \mathbf{0}))}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\right\}.$$

4.3 Frailty Model

It is assumed for Cox's proportional hazards model that the survival time of subjects are independent. This assumption might be violated in some situations. For example, the subjects share a common genetic makeup as they come from the same family, or share a common unmeasured environment as there might be repeated measurements on the same patient. In these cases, it is not reasonable to assume that the collected data are independent. Marginal model ignoring the dependency of observations yields a consistent estimate for the interested parameters, but it suffers its inefficiency when there are a large number of members in subgroups (See Lee *et al* , 1992 and also Table 4.3 below). One popular approach to model correlated survival time is to use a frailty model. A frailty corresponds to the random block effect that acts multiplicatively on the hazard rates of all subgroup subjects. In this section, we only consider Cox' proportional hazard frailty model, which is the most commonly used model for frailty. Here it is assumed that the hazard rate for the j^{th} subject in the i^{th} subgroup is

$$h_{ij}(t|\mathbf{x}_{ij}, u_i) = h_0(t) u_i \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}), \;\; i = 1, \cdots, G, j = 1, \cdots, n_i, \tag{4.3.1}$$

where the u_i's are associated with frailties, which are an independent and identically distributed sample from some population. It is common to assume that given the frailty u_i, the observations in the i-th group are independent. The most frequently used distribution for frailty is gamma distribution due to its mathematical simplicity. It is often assumed that the mean of frailty is 1 so that all parameters involved are estimable. Thus the density function of the frailty in gamma frailty model is given by

$$g(u) = \frac{\alpha^\alpha u^{\alpha-1} \exp(-\alpha\, u)}{\Gamma(\alpha)}.$$

From (4.2.1), the full likelihood of the data $\{(u_i, \mathbf{X}_{ij}, Z_{ij}, \delta_{ij}) : i = 1, \cdots, G, j = 1 \cdots, n_i\}$ we would have if the frailties were observed is given by

$$\prod_{i=1}^{G} \prod_{j=1}^{n_i} [\{h(z_{ij})|\mathbf{x}_{ij}, u_i)\}^{\delta_{ij}} \overline{F}(z_{ij}|\mathbf{x}_{ij}, u_i)] \prod_{i=1}^{G} g(u_i).$$

Integrating the full likelihood function with respect to $u_1, \cdots, u_G$, the likelihood of the observed data is given by

$$L = \exp\{\boldsymbol{\beta}^T (\sum_{i=1}^{G} \sum_{j=1}^{n_i} \delta_{ij} \mathbf{x}_{ij})\} \prod_{i=1}^{G} \frac{\alpha^\alpha \prod_{j=1}^{n_i} \{h_0(z_{ij})\}^{\delta_{ij}}}{\Gamma(\alpha) \{\sum_{j=1}^{n_i} H_0(z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha\}^{(A_i + \alpha)}},$$

where $A_i = \sum_{j=1}^{n_i} \delta_{ij}$. Therefore the penalized likelihood of the observed data is

$$\begin{aligned} &\sum_{i=1}^{G} \{\sum_{j=1}^{n_i} \delta_{ij} \log h(z_{ij}) - [(A_i + \alpha) \log \{\sum_{j=1}^{n_i} H_0(z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha\}]\} \\ &+ \sum_{i=1}^{G} \{\boldsymbol{\beta}^T (\sum_{j=1}^{n_i} \delta_{ij} \mathbf{x}_{ij}) + \alpha \log \alpha - \log \Gamma(\alpha)\} - \sum_{j=1}^{d} p_\lambda(|\beta_j|). \end{aligned}$$

Following Breslow's idea, consider the "least informative" nonparametric modeling for the $H_0(\cdot)$ ignoring the group effects, i.e.

$$H_0(z) = \sum_{l=1}^{N} \lambda_l I(z_l \leq z), \tag{4.3.2}$$

where $\{z_1, \cdots, z_N\}$ are the pooled observed failure times ignoring the random block effects.

Substitute (4.3.2) into the penalized likelihood of the observed data, then differentiating it with respect to λ_l, $l = 1, \cdots, N$, the root of the corresponding score function should satisfy the following equations:

$$\lambda_l^{-1} = \sum_{i=1}^{G} \frac{(A_i + \alpha) \sum_{j=1}^{n_i} I(z_l \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})}{\sum_{k=1}^{N} \lambda_k \sum_{j=1}^{n_i} I(z_k \leq z_{ij}) \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}) + \alpha}$$

for $l = 1, \cdots, N$.

With initial values for the $\alpha, \boldsymbol{\beta}$ and λ_l, it might be treated that the $h_0(z)$ and $H_0(z)$ are known in the likelihood function. Therefore the penalized likelihood estimate is the maximizer of the penalized likelihood of the observed data with respect to $(\alpha, \boldsymbol{\beta})$. When Newton-Raphson algorithm is applied for the penalized likelihood, it involves to compute the first two order derivatives of the gamma function, which causes some computational difficulties. One approach to avoid these difficulties is to maximize with respect to α over a grid of possible values for

the parameter α as suggested by Nielsen *et al* (1992), rather than over $[0, \infty)$. Our simulation experience shows that the estimate of $\boldsymbol{\beta}$ is quite empirically robust to the chosen grid of possible values for α.

A natural initial estimator for $\boldsymbol{\beta}$ is the pseudo-likelihood estimator of $\boldsymbol{\beta}$ in the marginal model. The corresponding $h_1, \cdots, h_N$ defined in the last section may be served as an initial estimator for $\lambda_1, \cdots, \lambda_N$. Hence given a value of α and initial values of $\boldsymbol{\beta}$ and $\lambda_1, \cdots, \lambda_N$, update the values of $\lambda_1, \cdots, \lambda_N$ and $\boldsymbol{\beta}$ in turn until the values of $\boldsymbol{\beta}$ and $\lambda_1, \cdots, \lambda_N$ converge. The proposed algorithm avoids to optimize a high-dimensional problem. It will give us an efficient estimate for $\boldsymbol{\beta}$. The algorithm may converge slowly or even not converge. In this situation, the idea of one-step estimator (see Bickel, 1973) provides us an alternative approach. See Chapter 2 for some other variation.

4.4 Simulation Studies and Applications

4.4.1 Local quadratic approximations and standard errors

As suggested in Chapter 2, one may use a local quadratic approximation for penalty functions in order to apply Newton-Raphson algorithm for the penalized partial likelihood function. Suppose that there is an initial value $\boldsymbol{\beta}_0$ is close to the maximizer of (4.2.5). When β_{j0} is not very close to 0, the penalty $p_\lambda(|\beta_j|)$ can be locally approximated by the quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta|)\text{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_{j0})/|\beta_{j0}|\}\beta_j,$$

otherwise, set $\hat{\beta}_j = 0$. In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}p'_\lambda(|\beta_{j0}|)(\beta_j^2 - \beta_{j0}^2), \quad \text{for} \quad \beta_j \approx \beta_{j0}.$$

Thus the maximization problems (4.2.5) can be reduced to a quadratic maximization problem and the Newton-Raphson algorithm can be used.

The standard errors for estimated parameters can be directly obtained because we are estimating parameters and selecting variables at the same time. Following the conventional technique in the likelihood setting, the corresponding sandwich formula can be used as an estimator for the covariance of the estimates $\hat{\boldsymbol{\beta}}$. For Cox's proportional hazards model, the solution in Newton-Raphson algorithm is updated by

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 - \{\nabla^2\ell(\boldsymbol{\beta}_0) - n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\{\nabla\ell(\boldsymbol{\beta}_0) - n\mathbf{U}_\lambda(\boldsymbol{\beta}_0)\}, \tag{4.4.1}$$

where $\ell(\boldsymbol{\beta})$ is the partial likelihood,

$$\nabla \ell(\boldsymbol{\beta}_0) = \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}, \quad \nabla^2 \ell(\boldsymbol{\beta}_0) = \frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$$

$$\Sigma_\lambda(\boldsymbol{\beta}_0) = \text{diag}\{p'_\lambda(|\beta_{10}|)/|\beta_{10}|, \cdots, p'_\lambda(|\beta_{d0}|)/|\beta_{d0}|\}, \quad \text{and} \quad \mathbf{U}_\lambda(\boldsymbol{\beta}_0) = \Sigma_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0.$$

Thus the corresponding sandwich formula is given by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1} \widehat{\text{cov}}\{\nabla \ell(\hat{\boldsymbol{\beta}})\}\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1}. \tag{4.4.2}$$

This formula will be shown to have good accuracy for moderate sample sizes. The sandwich formula for the frailty model can be derived by the same way.

4.4.2 Selection of thresholding parameters

To implement the methods described in previous sections, it is desirable to have an automatic method for selecting the thresholding parameter λ involved in $p_\lambda(\cdot)$ based on the data. Such a procedure is analogous to an automatic subset selection procedure such as forward selection, backward elimination and all subset regression. Here we estimate λ via minimizing an approximate generalized cross-validation (GCV) statistic (Craven and Wahba, 1977). By some straightforward calculation, the effective number of parameters for Cox's proportional hazards model in Newton-Raphson algorithm iteration is

$$e(\lambda) = \text{tr}[\{\nabla^2 \ell(\hat{\boldsymbol{\beta}}) - \Sigma_\lambda(\hat{\boldsymbol{\beta}})\}^{-1} \nabla^2 \ell(\hat{\boldsymbol{\beta}})].$$

Therefore the generalized cross-validation statistic is

$$\text{GCV}(\lambda) = \frac{-\ell(\hat{\boldsymbol{\beta}})}{n\{1 - e(\lambda)/n\}^2}$$

and $\hat{\lambda} = \text{argmin}_\lambda\{\text{GCV}(\lambda)\}$. The corresponding generalized cross-validation statistic can be derived for frailty model via the same way.

4.4.3 Prediction and model error

As discussed in Chapter 2, when the covariate $\mathbf{x}$ is random, if $\hat{\mu}(\mathbf{x})$ is a prediction procedure constructed using the present data, the prediction error is defined as

$$\text{PE}(\hat{\mu}) = E\{Y - \hat{\mu}(\mathbf{x})\}^2,$$

where the expectation is only taken with respect to the new observation $(\mathbf{x}, Y)$. The prediction error can be decomposed as

$$\mathrm{PE}(\hat{\mu}) = \{Y - E(Y|\mathbf{x})\}^2 + E\{E(Y|\mathbf{x}) - \hat{\mu}(\mathbf{x})\}^2.$$

The first component is inherent by due to stochastic errors. The second component is due to lack of fit to an underlying model. This component is called model error and is denoted by $\mathrm{ME}(\hat{\mu})$. For Cox's proportional hazards model,

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}).$$

Therefore

$$\mu(\mathbf{x}) = E(T|\mathbf{x}) = \int_0^\infty t\, h_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta})\exp\{-\int_0^t h_0(u)\exp(\mathbf{x}^T\boldsymbol{\beta})\,du\}\,dt.$$

In the following simulation examples, it will be taken that $h_0(t) \equiv 1$. Thus by some algebraic calculation,

$$\mu(\mathbf{x}) = \exp(-\mathbf{x}^T\boldsymbol{\beta}).$$

For Cox's proportional hazards frailty model with $h_0(t) \equiv 1$ as in the simulation example below,

$$\mu(\mathbf{x}) = \exp(-\mathbf{x}^T\boldsymbol{\beta})E(u^{-1}).$$

The factor due to frailty can be ignored when we compare the performance of two different approaches in terms of their Relative Model Errors, denoted by RME.

4.4.4 Simulations

In the following examples, we numerically compare the proposed variable selection methods with the maximum (partial) likelihood estimate and best subset variable selection. All simulations are conducted using MATLAB codes. To find the best subset variable selection, we searched exhaustively over all possible subsets and selected a subset with the best BIC score.

Example 4.4.1 In this example we simulated 100 data sets consisting of $n = 50,\ 75$ and 100 observations from the exponential hazard model

$$h(t|\mathbf{x}) = \exp(\mathbf{x}^T\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$. The x_i were marginally standard normal and the correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = 0.5$. The distribution of the censoring time

is an exponential distribution with mean $U\exp(\mathbf{x}^T\boldsymbol{\beta}_0)$, where U is randomly generated from the uniform distribution over [1,3] for each simulated data set so that about 30% data are censored, and $\boldsymbol{\beta}_0 = \boldsymbol{\beta}$ which is regarded as a known constant such that the censoring scheme is noninformative. This model will give us that when the sample size n equals to 50, the standard error of the maximum partial likelihood estimator $\hat{\beta}_7$ is about 0.3. Thus the true signal 0.6 is about twice standard deviation of its estimator. The intention of this choice is to show how the best subset variable selection and the penalized likelihood estimate with hard-thresholding penalty, referred to as HARD, are affected by the stochastic errors inherited in variable selection process.

Model errors of the proposed procedures are compared to those of the maximum partial likelihood estimates. The Median of Relative Model Errors (MRME) over 100 simulated data sets are summarized in Table 4.1. The average of 0 coefficients is also reported in Table 2.3, in which the column labeled "correct" presents the average restricted only to the true zero coefficients, while the columns labeled "incorrect" depicts the average of coefficients erroneously set to 0. From Table 4.1, it can be seen that when the sample size is relatively small, LASSO performs the best in terms of model error and reduction of model complexity. However, when the sample size increases, the performance of LASSO becomes worse, as it creates some excessive modeling bias. When the sample size is moderate size, SCAD outperforms the other three methods. Compared with SCAD, the performance of the best subset variable selection is worse than that of SCAD in terms of MRME, as it ignores the stochastic errors inherited in variable selection process. The same conclusion can be drawn for HARD.

The accuracy of the proposed standard error formula is illustrated. The median absolute deviation divided by 0.6745, denoted by SD in Table 4.2, of 100 estimated coefficients in the 100 simulations can be regarded as the true standard error. The median of the 100 estimated SDs, denoted by SD_m, and the median absolute deviation error of 100 estimated standard errors divided by 0.6745, denoted by SD_{mad}, gauge the overall performance of the standard error formula. Table 4.2 only presents the results for non-zero coefficients when sample size $n = 100$. The results for the other two cases with $n = 50$ and 75 are similar. Table 4.2 suggests that the proposed standard error formula performs well.

Example 4.4.2 In this example we simulated 100 data sets consisting of G groups and n subjects in each group from the exponential hazard frailty model

$$h(t|\mathbf{x}, u) = u\exp(\mathbf{x}^T\boldsymbol{\beta}),$$

Table 4.1: Simulation results for Cox's proportional hazards model

Method	MRME(%)	Aver. no. of 0 Coeff.	
		correct	incorrect
$n = 50$			
SCAD	0.2682	4.40	0.35
LASSO	0.1273	4.10	0.21
HARD	0.2962	4.76	0.64
Best subset	0.3164	4.56	0.33
$n = 75$			
SCAD	0.3696	4.34	0.18
LASSO	0.4559	4.05	0.10
HARD	0.3866	4.87	0.29
Best subset	0.4505	4.75	0.15
$n = 100$			
SCAD	0.3346	4.33	0.03
LASSO	0.4582	3.99	0.02
HARD	0.4367	4.84	0.08
Best subset	0.4624	4.76	0.06

Table 4.2: Standard deviations for Cox's proportional hazards models ($n = 100$)

	$\hat{\beta}_1$		$\hat{\beta}_4$		$\hat{\beta}_7$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.191	0.154 (0.023)	0.143	0.169 (0.020)	0.188	0.141 (0.026)
LASSO	0.190	0.150 (0.019)	0.163	0.143 (0.016)	0.157	0.118 (0.016)
HARD	0.180	0.161 (0.019)	0.155	0.175 (0.014)	0.144	0.147 (0.015)
Best subset	0.184	0.161 (0.019)	0.154	0.175 (0.015)	0.141	0.148 (0.015)

where $\boldsymbol{\beta}$ and $\mathbf{x}$ are the same as those in Example 4.2.1, and the frailty u is distributed according to a standard exponential distribution. First we compare the performance of the proposed approach to estimating the regression coefficients and the pseudo-likelihood approach. Here we compare their MRMEs and accuracy in terms of standard deviations of estimated coefficients. To save space, Table 4.3 presents their MRMEs and only the standard deviations of $\hat{\beta}_1$. Results for other estimated coefficients are almost the same. The notation in Table 4.3 is the same as that in Table 4.2. Compared the true standard deviations with the estimated standard deviations of the proposed approach, it can be seen that from Table 4.3 that the proposed sandwich formula performs surprising well. From Table 4.3, it can be found that when the number of members in each group is small, such as $n = 2$, the pseudo-likelihood approach performs better

than the proposed approach. This phenomenon is consistent with the conclusion of Lee *et al* (1992). Intuitively, the dependency among the samples may be ignored when the number of each group is small. In this situation the proposed approach introduces a number of nuisance parameters which create some extra variation, compared with the pseudo-likelihood approach. However, when the number of members in each group is relatively large, the proposed approach outperforms the pseudo-likelihood approach. This suggests that the pseudo likelihood estimate may result in a poor model when the dependency among the samples cannot be ignored.

Table 4.3: Comparisons between the proposed approach and the pseudo likelihood approach

	proposed method		pseudo likelihood		MRME
(G,n)	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	(Pseudo/proposed)
(50,2)	0.270	0.217 (0.039)	0.162	0.173 (0.029)	0.4051
(75,2)	0.219	0.173 (0.028)	0.139	0.133 (0.013)	0.5883
(100,2)	0.192	0.151 (0.017)	0.095	0.111 (0.010)	0.9234
(20,10)	0.114	0.106 (0.018)	0.135	0.114 (0.015)	1.9139
(10,20)	0.111	0.088 (0.023)	0.154	0.113 (0.014)	3.7974
(20,20)	0.076	0.067 (0.011)	0.081	0.075 (0.006)	7.3268
(10,40)	0.085	0.067 (0.018)	0.118	0.076 (0.007)	6.4120

Now the performance of variable selection via nonconcave penalized likelihood and best subset variable selection are compared in terms of their model errors, reduction of model complexity and accuracy. Model errors of the proposed procedures are compared to those of the maximum likelihood estimates. The Median of Relative Model Errors (MRME) over 100 simulated data sets with $G = 50,\ 75$ and 100 groups and 2 members in each group are summarized in Table 4.4, and standard errors for estimated nonzero coefficients with $G = 100$ and 2 members in each group are depicted in Table 4.5. From Tables 4.4 and 4.5, one may draw the same conclusion as those in Example 4.4.1.

4.4.5 Application

The proposed approach will be applied to the "nursing home" data set analyzed by Morris, Norton and Zhou (1994), which gives a full description of this data set. Here is a brief description of this data set. The data were from an experiment sponsored by the National Center for Health Services Research in 1980-1982, involving 36 for-profit nursing homes in San Diego, California. The experiment was designed to assess the effects of differing financial incentives on the admission of nursing home patients, on their subsequent care, and on the durations of

Table 4.4: Simulation results for frailty model

Method	MRME(%)	Aver. no. of 0 Coeff.	
		correct	incorrect
$G = 50,\ n = 2$			
SCAD	0.5962	4.65	0.34
LASSO	0.4249	4.14	0.12
Hard	0.6694	4.66	0.33
Best subset	0.7754	4.69	0.23
$G = 75,\ n = 2$			
SCAD	0.5721	4.64	0.11
LASSO	0.5332	4.14	0
Hard	0.6218	4.74	0.10
Best Subset	0.5879	4.81	0.07
$G = 100,\ n = 2$			
SCAD	0.5781	4.43	0
LASSO	0.7436	4.18	0
Hard	0.6203	4.72	0.05
Best subset	0.6708	4.84	0.01

Table 4.5: Standard deviations for frailty model ($G = 100,\ n = 2$)

	$\hat{\beta}_1$		$\hat{\beta}_4$		$\hat{\beta}_7$	
Method	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$	SD	$SD_m(SD_{mad})$
SCAD	0.189	0.148 (0.020)	0.174	0.122 (0.017)	0.088	0.126 (0.021)
LASSO	0.188	0.121 (0.015)	0.175	0.099 (0.012)	0.133	0.084 (0.021)
Hard	0.180	0.147 (0.019)	0.178	0.133 (0.022)	0.132	0.129 (0.021)
Best subset	0.173	0.149 (0.018)	0.170	0.134 (0.021)	0.110	0.130 (0.020)

stay. The 18 treatment nursing homes received higher per diem payments for accepting more disabled medicaid patients. They also received bonuses for improving a patient's health status and for discharging patients to their homes within 90 days. These incentives were not offered to the 18 control nursing homes. The total of 1601 samples are available.

Morris *et al* (1994) took *days in the nursing home* as the response variable t. They suggested the use of the following model:

$$h(t|\mathbf{x}) = h_0(t)\exp(\sum_{i=0}^{7} x_i\beta_i), \tag{4.4.3}$$

where x_1 is a treatment indicator, being 1 if treatment nursing home and 0 if control; x_2 is

variable *age*, whose range is from 65 to 90. x_3 is variable *gender*, being 1 if male and 0 if female; x_4 is a marital status indicator, being 1 if married and 0 otherwise; x_5 , x_6 and x_7 are three binary health status indicators, corresponding the best health to the worst health. β_0 is an intercept when a parametric model for the baseline h_0 is employed, while it is dropped from the model if Cox's proportional hazard model is applied such that the model is identifiable. Morris *et al* (1994) applied three parametric models and the Cox regression model for this data set. It is clear that the proposed model does not include any interactions. To illustrate the proposed approach, all interactions among treatment, age, gender and marital status are included in the initial model, and fit the data by Cox's regression model with 13 covariates. Only x_2 is standardized as other variables are binary. Penalized partial likelihood approach with the SCAD, L_1 and hard penalty are applied for this data set. The thresholding parameter λ, selected by the GCV, are 0.0227, 0.0113 and 0.0890 for the SCAD, LASSO and HARD respectively. Best subset variable selection with AIC and BIC are also conducted. Estimated coefficients and their standard errors are shown in Table 4.6.

Table 4.6: Estimated coefficients and standard errors (Cox's model)

	MLE	Best (BIC)	Best (AIC)	SCAD	LASSO	hard
TRT	-0.04 (0.07)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)
Age	-0.12 (0.05)	0 (−)	-0.09 (0.03)	-0.09 (0.04)	-0.05 (0.02)	0 (−)
Male	0.43 (0.10)	0.40 (0.06)	0.44 (0.08)	0.44 (0.08)	0.31 (0.05)	0.44 (0.08)
Married	0.22 (0.14)	0 (−)	0.16 (0.08)	0.18 (0.08)	0.08 (0.03)	0.18 (0.08)
Health1	0.03 (0.08)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)
Health2	0.24 (0.07)	0.24 (0.06)	0.23 (0.06)	0.23 (0.06)	0.14 (0.04)	0.23 (0.06)
Health3	0.57 (0.10)	0.53 (0.09)	0.54 (0.09)	0.54 (0.09)	0.35 (0.06)	0.55 (0.09)
TRT*Age	0.07 (0.06)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)
TRT*Male	-0.10 (0.13)	0 (−)	-0.15 (0.11)	-0.16 (0.11)	0 (−)	-0.15 (0.11)
TRT*Married	-0.00 (0.16)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)
Age*Male	0.16 (0.06)	0 (−)	0.17 (0.06)	0.16 (0.06)	0.07 (0.03)	0.09 (0.05)
Age*Married	0.09 (0.08)	0 (−)	0 (−)	0.09 (0.08)	0 (−)	0 (−)
Male*Married	-0.07 (0.16)	0 (−)	0 (−)	0 (−)	0 (−)	0 (−)

From Table 4.6, the best subset variable selection with AIC and the SCAD yield almost the same model. Compared with other approaches, LASSO somewhat shrinks all nonzero coefficients, while best subset variable selection with BIC results in too simple model as it over-penalizes dimension of selected model, compared with AIC, although the performance of AIC and BIC is similar when sample size is small.

Table 4.7: Estimated coefficients and standard errors (frailty model)

Method	MLE	Best (BIC)	Best (AIC)	SCAD	LASSO	hard
TRT	-0.05 (0.05)	0(−)	0(−)	0(−)	0(−)	0(−)
Age	-0.11 (0.02)	0(−)	-0.08 (0.02)	-0.08 (0.02)	-0.07 (0.02)	-0.09 (0.02)
Male	0.40 (0.08)	0.39 (0.04)	0.36 (0.04)	0.36 (0.03)	0.32 (0.04)	0.35 (0.04)
Married	0.25 (0.12)	0(−)	0.18 (0.07)	0.18 (0.07)	0.14 (0.05)	0.21 (0.08)
Health1	0.09 (0.07)	0(−)	0(−)	0.09 (0.07)	0(−)	0.09 (0.07)
Health2	0.27 (0.07)	0.24 (0.07)	0.24 (0.08)	0.27 (0.06)	0.19 (0.06)	0.27 (0.07)
Health3	0.58 (0.07)	0.54 (0.08)	0.54 (0.07)	0.58 (0.07)	0.45 (0.07)	0.58 (0.07)
TRT*Age	0.05 (0.04)	0(−)	0(−)	0(−)	0.01 (0.01)	0(−)
TRT*Male	-0.07 (0.07)	0(−)	0(−)	0(−)	0(−)	0(−)
TRT*Married	-0.04 (0.14)	0(−)	0(−)	0(−)	0(−)	0(−)
Age*Male	0.17 (0.06)	0(−)	0.18 (0.06)	0.18 (0.06)	0.12 (0.05)	0.17 (0.06)
Age*Married	0.08 (0.07)	0(−)	0(−)	0(−)	0(−)	0.07 (0.06)
Male*Married	-0.05 (0.15)	0(−)	0(−)	0(−)	0(−)	0(−)

The resulting model is somewhat not consistent with the one without including interactions, analyzed by Morris *et al* (1994). The difference between the models in Morris *et al* (1994) and the resulting models here is summarized as follows.

When model excluding interactions, the *age* variable is not statistically significant, pointed out by Morris *et al* (1994). However, it is very significant in the resulting model with interactions. It is clear from Table 4.6 that elderly patients are more likely stay at nursing home.

From Table 4.6, the interactions between the variables *treatment* and *gender*, and *age* and *gender* are significant, although the variable *treatment* is not significant, as pointed out by Morris *et al* (1994). It seems that men prefer to stay nursing home with treatment, while elderly men like to leave nursing home earlier. The latter is because elderly men are much more likely to be married (see Morris *et al* , 1994), and they like to stay at their own home rather than nursing home.

The data were collected from 36 nursing homes. Each nursing home might have some unmeasured environment, therefore patients in each nursing home might share common frailty. To address this issue, a semi-parametric gamma frailty model is applied to analyze this data set. The unknown parameter α in gamma frailty was estimated by searching over 20 log-space grids on the interval of [0.01,10]. The resulting estimator $\hat{\alpha}$ is 0.2041 for the proposed

approaches. Thresholding parameter for SCAD, LASSO and HARD are 0.0207, 0.0059 and 0.0703, respectively. Estimated coefficients and standard errors are reported in Table 4.7.

Comparing results in Table 4.6 and 4.7, Cox's proportional hazards model and Cox's frailty model yield almost the same results in terms of their estimated regression coefficients, while the accuracy of the estimated coefficients was improved by using Cox's frailty model.

It is worth to noting that although best subset variable selection with AIC yields similar model to those obtained via nonconcave penalized likelihood approaches, best subset variable selection is much more expensive in computation. The proposed penalized likelihood approach with the proposed algorithm can save computation time in an order of hundreds in this example, compared with naive implementation of best subset variable selection with exhaustively searching all possible subsets.

4.5 Proofs

We first introduce some notation. Let

$$\widehat{P}(z) = \frac{1}{n}\sum_{i=1}^{n} \delta_i I(Z_i \leq z) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \leq z, \delta = 1)$$

and

$$P(z) = P(Z \leq z, \delta = 1).$$

Rewrite the log-partial likelihood function $\ell(\boldsymbol{\beta})$ as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{x}_i^T - \sum_{i=1}^{n} [\delta_i \log\{\sum_{j \in R_i} \exp(\mathbf{x}_j^T \boldsymbol{\beta})\}].$$

Therefore the log-partial likelihood function $\ell(\boldsymbol{\beta})$ can be represented as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{x}_i^T \boldsymbol{\beta} - n \int \log \sum_{j=1}^{n} \exp(\mathbf{x}_j^T \boldsymbol{\beta}) I(Z_j \geq z) d\, \widehat{P}(z).$$

Denote

$$A(\boldsymbol{\beta}) = \log \sum_{j=1}^{n} \exp(\mathbf{x}_j^T \boldsymbol{\beta}) I(Z_j \geq z).$$

Therefore

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \mathbf{x}_i^T \boldsymbol{\beta} - n \int A(\boldsymbol{\beta}) \widehat{P}(z).$$

Proof of Theorem 4.1:

Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) < Q(\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon. \tag{4.5.1}$$

This implies with probability at least $1 - \varepsilon$ that there exists a local maximum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \mathbf{u} : \|\mathbf{u}\| \leq C\}$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\alpha_n)$.

Using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D_n(\mathbf{u}) &\equiv Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - Q(\boldsymbol{\beta}_0) \\ &\leq \ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0) - n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}, \end{aligned}$$

where s is the number of components of $\boldsymbol{\beta}_{10}$. Therefore

$$\begin{aligned} D_n(\mathbf{u}) &\leq \sqrt{n}\alpha_n \mathbf{u}^T \left(\frac{1}{\sqrt{n}}\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right) - \frac{1}{2} n\alpha_n^2 \mathbf{u}^T \left\{\int \frac{\partial^2 A(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} d\widehat{P}(z)\right\} \mathbf{u} \\ &\quad - n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}, \end{aligned} \tag{4.5.2}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}$ and $\boldsymbol{\beta}_0$.

Note that by the standard argument in the theory of partial likelihood (see for example Tsiatis 1981 and Wong, 1986), it follows that $n^{-1/2}\frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} = O_p(1)$, thus the first term in the right-hand side of (4.5.2) is of the order $O_P(n^{1/2}\alpha_n) = O_P(n\alpha_n^2)$. It is easy to show that $n\sum_{j=1}^{s}\{p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)\}$ is bounded by

$$\sqrt{s}\sqrt{n}\alpha_n a_n \|\mathbf{u}\| + n\alpha_n^2 b_n \|\mathbf{u}\|^2.$$

To deal with the second term, let us simplify it first:

$$\begin{aligned} \frac{\partial^2 A(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} &= \frac{\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^T \exp(\mathbf{x}_j^T \boldsymbol{\beta}^*) I(Z_j \geq z)}{\sum_{j=1}^{n}\{\exp(\mathbf{x}_j^T \boldsymbol{\beta}^*) I(Z_j \geq z)\}} \\ &\quad - \frac{\{\sum_{j=1}^{n} \mathbf{x}_j \exp(\mathbf{x}_j^T \boldsymbol{\beta}^*) I(Z_j \geq z)\}\{\sum_{j=1}^{n} \mathbf{x}_j \exp(\mathbf{x}_j^T \boldsymbol{\beta}^*) I(Z_j \geq z)\}^T}{[\sum_{j=1}^{n}\{\exp(\mathbf{x}_j^T \boldsymbol{\beta}^*) I(Z_j \geq z)\}]^2} \\ &= \left\{\frac{E\{\mathbf{x}\mathbf{x}^T \exp(\mathbf{x}^T \boldsymbol{\beta}_0) I(Z \geq z)\}}{E\{\exp(\mathbf{x}^T \boldsymbol{\beta}_0) I(Z \geq z)\}}\right. \\ &\quad \left. - \frac{\{E\mathbf{x} \exp(\mathbf{x}^T \boldsymbol{\beta}_0) I(Z \geq z)\}\{E\mathbf{x} \exp(\mathbf{x}^T \boldsymbol{\beta}_0) I(Z \geq z)\}^T}{[E\{\exp(\mathbf{x}^T \boldsymbol{\beta}_0) I(Z \geq z)\}]^2}\right\}\{1 + o_P(1)\} \\ &\equiv I_0(z; \boldsymbol{\beta}_0)\{1 + o_P(1)\} \end{aligned}$$

as $E\{\exp(2\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}\mathbf{x}^T\}$ is bounded uniformly in the neighborhood of $\boldsymbol{\beta}_0$ by assumption.

Note that $\{\int I_0(z;\boldsymbol{\beta}_0)\,dP(z)\}^{-1}$ is the covariance matrix of maximum partial likelihood estimator (see Tsiatis, 1981), thus it is finite and positive definite. By assumption that $b_n \to 0$, the second term of (4.5.1) will dominate the other two terms uniformly in $\|\mathbf{u}\| = C$, by choosing a sufficiently large C. Hence (4.5.1) holds. This completes the proof of Theorem 4.1.

Proof of Lemma 4.1:

It is sufficient to show that with probability tending to 1 as $n \to \infty$, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_P(n^{-1/2})$ and for some small $\varepsilon_n = Cn^{-1/2}$ and $j = s+1, \cdots, d$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for} \quad 0 < \beta_j < \varepsilon_n, \tag{4.5.3}$$

and

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} > 0 \quad \text{for} \quad -\varepsilon_n < \beta_j < 0. \tag{4.5.4}$$

To show (4.5.3), by Taylor's expansion, we have

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} - np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) \\
&= \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 \ell(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{l0}) - np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)
\end{aligned}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that by the same arguments as the proof of Theorem 4.1, the order of the first term is $O_P(n^{-1/2})$. By the assumption that $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_P(n^{-1/2})$, the order of the second term is also $O_P(n^{-1/2})$.Thus

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n\{-\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_P(n^{-1/2}/\lambda_n)\}.$$

Since $\lim_{n\to\infty} \liminf_{\theta\to 0^+} \lambda_n^{-1}p'_{\lambda_n}(\theta) > 0$ and $n^{-1/2}/\lambda_n \to 0$, the sign of the derivative is completely determined by that of β_j. Hence, (4.5.3) and (4.5.4) follow. This completes the proof.

Proof of Theorem 4.2:

It follows by Lemma 4.1 that Part (i) holds. Now we prove Part (ii). It can be shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 4.1 that is a root n consistent local maximizer of $Q\{\begin{pmatrix}\boldsymbol{\beta}_1 \\ \mathbf{0}\end{pmatrix}\}$,

regarded as a function of $\boldsymbol{\beta}_1$, and satisfying the likelihood equations:

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}} = 0, \quad \text{for} \quad j = 1, \cdots, s. \tag{4.5.5}$$

Note that $\hat{\boldsymbol{\beta}}_1$ is a consistent estimator,

$$\begin{aligned} & \left.\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}\right|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_1\\ \mathbf{0}\end{pmatrix}} - np'_{\lambda_n}(|\hat{\beta}_j|) \\ = \; & \frac{\partial \ell(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s}\left\{\frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} + o_P(1)\right\}(\hat{\beta}_l - \beta_{l0}) \\ & -n\left(p'_{\lambda_n}(|\beta_j^0|)\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_n}(|\beta_{j0}|) + o_P(1)\}(\hat{\beta}_j - \beta_{j0})\right). \end{aligned}$$

Note that by standard arguments in the theory of partial likelihood (see, for example, Tsiatis, 1981 and Wong, 1986), and by Slutsky's Theorem, it follows that

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

in distribution, where

$$I_1(\boldsymbol{\beta}_{10}) \equiv I(\boldsymbol{\beta}_{10}, \mathbf{0}) = -E\left\{\frac{\partial^2 \ell(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_1 \partial \boldsymbol{\beta}_1^T}\right\}.$$

It can be shown that

$$I_1(\boldsymbol{\beta}_{10}) = \int I_{10}(z; (\boldsymbol{\beta}_{10}, \mathbf{0}))\, dP(z). \tag{4.5.6}$$

where $I_{10}(z; (\boldsymbol{\beta}_{10}, \mathbf{0}))$ consists of the first s rows and columns of $I_0(z; (\boldsymbol{\beta}_{10}, \mathbf{0}))$. See Tsiatis (1981) for details.

4.6 Conclusion

Variable selection via nonconcave penalized likelihood has been successfully extended to Cox's proportional hazards model and frailty model in this chapter. Consistency and oracle properties has been established for the proposed estimators. Numerical comparisons were conducted. From the numerical comparison, it may be concluded that the SCAD has good theoretic properties and excellent performance in terms of reduction of model complexity and model error. Further the SCAD retains good features of LASSO and best subset variable selection.

Chapter 5

Introduction and Literature Review for Part II

Kernel smoothing methods for curve estimation are useful tools for data analysis. There is a huge literature on this subject. Some excellent reference books include Härdle (1990), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Bowman and Azzalini (1997), Hart (1997) and Loader (1999). See these monographs for access to the literature. This chapter gives a brief summary of kernel regression in the univariate case in order to provide some backgrounds of local modeling. The ideas of Sections 5.1 and 5.2 are motivated from Sections 1.2, 2.2 and 2.3 of Fan and Gijbels (1996). Section 5.1 describes the conventional kernel regression estimators, including the Nadaraya-Watson (NW) kernel estimator and Gasser-Müller (GM) kernel estimator. Local polynomial regression and the local likelihood approach are introduced in Sections 5.2 and 5.3 respectively. Finally Section 5.4 presents the organization of the second part of this dissertation.

5.1 Conventional Kernel Regression

Consider the bivariate data $(X_1, Y_1), \cdots, (X_n, Y_n)$, an i.i.d. sample from the model:

$$Y = m(X) + \varepsilon,$$

where ε is random error with $E(\varepsilon|X) = 0$ and $\text{var}(\varepsilon|X = x) = \sigma^2(x)$. The nonparametric regression problem is to estimate the regression function $m(\cdot)$. The shape of $m(x)$ reveals

interesting features in the data. Usually a datum point closer to x carries more information about the value of $m(x)$, therefore an intuitive estimator for the regression function $m(x)$ is the running local average. An improved version of this is the locally weighted average. That is

$$\hat{m}(x) = \sum_{i=1}^{n} w_i(x) Y_i / \sum_{i=1}^{n} w_i(x).$$

An alternative interpretation of locally weighted average estimators is that the resulting estimator is the solution of the following weighted least-squares problem:

$$\min_{\theta} \sum_{i=1}^{n} (Y_i - \theta)^2 w_i.$$

In other words, the conventional kernel regression estimators are a weighted least squares estimate at the point x with a local constant approximation.

Both NW estimator and GM estimator are locally weighted average estimators. Let K be a real-valued function assigning weights. The function K is usually a symmetric probability density and is called a kernel function. Let h be a bandwidth, which is a nonnegative number controlling the size of the local neighborhood. Denote $K_h(\cdot) = K(\cdot/h)/h$.

Setting the weights $w_i(x) = K_h(X_i - x)$ results in the NW kernel regression estimator, which is given by

$$\hat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x) Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}. \tag{5.1.1}$$

See Nadaraya (1964) and Watson (1964).

Since the denominator in (5.1.1) is a random variable, it is inconvenient when deriving the asymptotic properties of the estimator. Assume that the data have already been sorted according to the X-variable. Taking the local weights $w_i(x) = \int_{s_{i-1}}^{s_i} K_h(u - x) dx$ with $s_i = (X_i + X_{i+1})/2$, $X_0 = -\infty$ and $X_{n+1} = +\infty$, we obtain the GM regression estimator given by

$$\hat{m}_h(x) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(u - x)\, du Y_i.$$

See Gasser and Müller (1984).

It is worth noting that the choice of bandwidth h is of crucial importance in nonparametric kernel smoothing regression. If h is chosen too large then the resulting estimate misses fine features of the regression curve, while if h is selected too small then spurious sharp structure becomes visible. The choice of kernel function is not an important issue in nonparametric

regression. Commonly used kernel functions include the Gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}),$$

and the *symmetric Beta family*:

$$K(t) = \frac{1}{Beta(1/2, \gamma)}(1 - t^2)_+^{\gamma}, \quad \gamma = 0,\ 1, \cdots, \tag{5.1.2}$$

where the subscript + denotes the positive part, which is assumed to be taken before exponentiation. The corresponding kernel functions when $\gamma = 0$, 1, 2 and 3 are respectively the uniform, the Epanechnikov, the biweight and the triweight kernel functions.

The basic asymptotic properties of the NW regression estimator and GM regression estimator have been well established. The asymptotic biases and variances of these two estimator are depicted in Table 5.1, taken from Fan (1992) and originate from the work of Mack and Müller (1989) and Chu and Marron (1991).

Table 5.1: Pointwise asymptotic biases and variances of smooth regression estimators

Method	Bias	Variance
NW estimator	$\{m''(x) + \frac{2m'(x)f'(x)}{f(x)}\}b_n$	V_n
GM estimator	$m''(x)b_n$	$1.5V_n$
Local linear	$m''(x)b_n$	V_n

Here $B_n = \frac{1}{2}\int_{-\infty}^{+\infty} u^2 K(u)\, du\, h^2$ and $V_n = \frac{\sigma^2(x)}{f(x)nh}\int_{-\infty}^{+\infty} K^2(u)\, du$.

5.2 Local Polynomial Regression

As indicated in the last section, both the NW estimator and the GM estimator are a local constant fit. It is natural to extend this to a local polynomial fit. The idea of local polynomial regression has been around for a long time. Since both a local constant and local polynomial fits use effectively datum points in a local neighborhood, this idea is referred as "local regression". It first appeared in the statistical literature in Stone (1977) and Cleveland (1979). Stone (1980, 1982) shows that local regression achieves optimal rates in a minimax sense. Müller (1987) establishes the equivalence between a local polynomial fit and a local constant fit under an equally-spaced design model. Fan (1992, 1993) focus on local linear regression in the random design case and show that it has many advantages, such as simple expression for local bias and

variance, spatial adaptation and high minimax efficiency. Fan and Gijbels (1992) proved that theoretically the local linear regression estimator adapts automatically to the boundary. This was also empirically observed by Tibshirani and Hastie (1987). Ruppert and Wand (1994) extended the results of Fan and Gijbels (1992) to the case of local polynomial estimation. A thorough study of this topic can also be found in Chapters 3 and 4 of Fan and Gijbels (1996).

Suppose that the regression function m is smooth. For z in a neighborhood of x, it follows by using Taylor's expansion that

$$m(z) \approx \sum_{j=1}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=1}^{p} \beta_j (z-x)^j. \tag{5.2.1}$$

Thus for X_i close enough to x,

$$m(X_i) \approx \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \equiv \mathbf{X}_i^T \boldsymbol{\beta},$$

where $\mathbf{X}_i = (1, (X_i - x_0), \cdots, (X_i - x_0)^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$. Intuitively datum points further from x have less information about $m(x)$. This suggests using a locally weighted polynomial regression

$$\sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 K_h(X_i - x), \tag{5.2.2}$$

where $K(\cdot)$ denotes a kernel function and h is a bandwidth. Denote by $\hat{\beta}_j$ $(j = 0, \cdots, p)$ the minimizer of (5.2.2). The above exposition suggests that an estimator for $m^{(\nu)}(x)$ is

$$\hat{m}_\nu(x) = \nu! \hat{\beta}_\nu. \tag{5.2.3}$$

For estimating the first derivative $m'(\cdot)$, one takes the slope of the local quadratic regression, a local cubic fit can be used for estimating the second derivative function, and so on. See for example Section 3.3 of Fan and Gijbels (1996). In general, local polynomial fitting has certain advantages over the NW and the GM estimator not only for estimating regression curves, but also for derivative estimation.

5.3 Local Likelihood Approach

The local likelihood approach was first proposed by Tibshirani and Hastie (1987) based on the running line smoother, whereas the methodology of Staniswalis (1989) relies on local constant fitting. As an extension of the local likelihood approach, local quasi-likelihood estimation

using local constant fits was considered by Severini and Staniwalis (1989). Fan, Heckman and Wand (1995) investigate the asymptotic properties of the local quasi-likelihood method using local polynomial modeling. Fan, Farmen and Gijbels (1998) addressed the issue of bandwidth selection, bias and variance assessment and constructed confidence intervals in local maximum likelihood estimation. Fan and Chen (1999) proposed one-step local quasi-likelihood estimation, and demonstrated that the one-step local quasi-likelihood estimator performs as well as the maximum local quasi-likelihood estimator using the ideal optimal bandwidth. Fan, Gijbels and King (1998) extended the idea of the local likelihood approach to local partial likelihood to the context of censored survival data analysis, such as Cox's regression model. The ideas in this section are motivated from Fan, Heckman and Wand (1995).

Generalized linear models introduced by Nelder and Wedderburn (1972) extend the traditional least squares fitting of linear models in a natural way. The relationship between a random variable with conditional distribution from an exponential family and a set of covariates is modeled by a linear fit to transformation of the conditonal mean. A comprehensive account of generalized linear models can be found in McMullagh and Nelder (1989). The definitions and notations of generalized linear models in this dissertation will follow those of McMullagh and Nelder (1989). Suppose that we have n independent observations $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ of random vector $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d-dimensional real vector of covariates, and Y is a scalar response variable. The conditional density of Y is a scalar response variable. The conditional density of Y given covariate $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family:

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\{[\theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}]/a(\phi) + c(y, \phi)\} \tag{5.3.1}$$

for known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$. In parametric generalized linear models it is usual to model a transformation of the regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ as linear, that is

$$\eta(\mathbf{x}) = g\{m(\mathbf{x})\} = \mathbf{x}^T\boldsymbol{\beta},$$

and g is a known *link* function. If $g = (b')^{-1}$, then g is called the canonical link because $b'\{\theta(\mathbf{x})\} = m(\mathbf{x})$.

Here are a few examples to illustrate the model (5.3.1). The first example is that the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ has a normal distribution with mean $m(\mathbf{x})$ and variance σ^2. The normal density can be rewritten as

$$f_{Y|\mathbf{X}} = \exp\left\{\frac{m(\mathbf{x})y - m^2(\mathbf{x})/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})\right\}.$$

It can be easily seen that

$$\phi = \sigma^2,\ a(\phi) = \phi,\ b(m) = m^2/2$$

and

$$c(y, \phi) = -y^2/(2\phi) - \log(\sqrt{2\pi\phi}).$$

The canonical link function is the identity link $g(t) = t$. This model is useful for a continuous response with homoscedastic errors.

Suppose that the conditional distributions of Y given $\mathbf{X} = \mathbf{x}$ is a Bernoulli distribution with success probability $p(\mathbf{x})$, in which case it can be seen that

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\left(y \log[p(\mathbf{x})/\{1 - p(\mathbf{x})\}] + \log\{1 - p(\mathbf{x})\}\right).$$

The canonical parameter in this example is $\theta(\mathbf{x}) = \text{logit}\{p(\mathbf{x})\}$, and the logit function is the canonical link.

Under model (5.3.1), it can be easily shown that the conditional mean and conditional variance are given respectively by $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = b'\{\theta(\mathbf{x})\}$, and $\text{var}(Y|\mathbf{X} = \mathbf{x}) = a(\phi)b''\{\theta(\mathbf{x})\}$. Since our primary interest is to estimate the mean function, without loss of generality, the factors related to the dispersion parameter ϕ are omitted. This leads to the following conditional log-likelihood function

$$\ell\{\theta, y\} = \theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}.$$

It has been of interest to adapt these models to situations where the functional form for the dependence of $g(m(\mathbf{x}))$ on $\mathbf{X}$ is unknown. In what follows, the covariate $\mathbf{X}$ is assumed to be a scalar random variable. If $\eta(x)$ is a smooth function of x, then for X_i close enough to an arbitrary point x_0,

$$\eta(X_i) \approx \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \equiv \mathbf{X}_i^T \boldsymbol{\beta}, \tag{5.3.2}$$

where $\mathbf{X}_i = (1, (X_i - x_0), \cdots, (X_i - x_0)^p)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$. Intuitively datum points close to x_0 have more information about $\eta(x_0)$ than data further from x_0, Therefore the local log-likelihood function based on the random sample $\{(X_i, Y_i)\}_{i=1}^n$ is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[Y_i (b')^{-1} \circ g^{-1}(\mathbf{X}_i^T \boldsymbol{\beta}) - b\{(b')^{-1} \circ g^{-1}(\mathbf{X}_i \boldsymbol{\beta})\} \right] K_h(X_i - x_0), \tag{5.3.3}$$

where $\circ$ denotes composition. Define the local maximum likelihood estimator of $\boldsymbol{\beta}$ to be

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta} \in R^{p+1}} \ell(\boldsymbol{\beta}).$$

Thus $\eta(x_0)$ and the ν-th derivative of $\eta(x_0)$ can be estimated by

$$\hat{\eta}(x_0) = \hat{\beta} \quad \text{and} \quad \hat{\eta}^{(\nu)} = \nu!\hat{\beta}_\nu$$

respectively, assuming η has p derivatives.

The choice of the link function g is not as crucial as for parametric generalized linear models, because the fitting is localized. Indeed it is conceivable to dispense with the link function and just estimate $m(x)$ directly. But there are several drawbacks to having the link equal to the identity. An identity link may yield a local likelihood that is not convex, allowing for the possibility of multiple maxima, inconsistency and computational problems. Use of the canonical link guarantees convexity. Furthermore the canonical link ensures that the final estimate has the correct range. For example, in the logistic regression context using the logit link leads to an estimate that is always a probability whereas using the identity link does not. A final reason for preferring the canonical link is that the estimate of $m(x)$ approaches the usual parametric estimate as the bandwidth becomes large. This can be useful as a diagnostic tool. See Fan, Heckman and Wand (1995) for details.

5.4 Organization of Part II

The second part of this dissertation contains Chapters 5, 6 and 7. Chapter 6 deals with statistical inference for varying-coefficient models. Local polynomial regression techniques are used to estimate coefficient functions and the asymptotic normality of the resulting estimators is established. The standard error formulas for estimated coefficients are derived and are empirically tested. A goodness-of-fit test technique, based on a nonparametric maximum likelihood ratio type of test, is also proposed to detect whether certain coefficient functions in a varying-coefficient model are constant or whether any covariates are statistically significant in the model. The null distribution of the test is estimated by a conditional bootstrap method. Our estimation techniques involve solving hundreds of local likelihood equations. To reduce computational burden, a one-step Newton-Raphson estimator is proposed and implemented. We show that the resulting one-step procedure can save computational cost in an order of tens without deteriorating its performance, both asymptotically and empirically. Both simulated and real data examples are used to illustrate our proposed methodology.

The SiZer map has been developed by Chaudhuri and Marron (1999) a methodology for finding which features in noisy data are strong enough to be distinguished from background

noise. It is based on scale space, i.e., a family of smooths of the data. However, their methods are inefficient for discrete data, such as binary data and count data. An extension of the SiZer map based on local likelihood is proposed in Chapter 7. Some computational implementation issues are addressed, and some applications for real data examples are discussed.

Chapter 6

Generalized Varying Coefficient Models

6.1 Introduction

Generalized linear models are widely used in many statistical applications. They are based on two fundamental assumptions: the conditional distributions belong to an exponential family and a known transform of the underlying regression function is linear. Various attempts have been made to relax the above model assumptions and hence widen their applicability, since a wrong model for the regression function can lead to excessive modeling biases and hence erroneous conclusions. The generalized varying-coefficient models (Hastie and Tibshirani 1993) widened the scope of applications by allowing regression coefficients to depend on certain covariates.

A motivation of this study comes from an analysis of an environmental data set, consisting of weekly measurements of pollutants and other environmental factors, collected in Hong Kong from January 1, 1994 to December 31, 1995 (Courtesy of Professor T. S. Lau). Of interest is the association between the levels of pollutants and the number of weekly total hospital admissions for circulationary and respiratory problems. It is natural to allow the association to change over time. Such a problem can be modeled as follows

$$g\{m(\mathbf{u},\, \mathbf{x})\} = \sum_{j=1}^{p} a_j(\mathbf{u})\, x_j \tag{6.1.1}$$

for some given link function $g(\cdot)$, where $\mathbf{x} = (x_1, \ldots, x_p)^T$, and $m(\mathbf{u},\, \mathbf{x})$ is the mean regression

function of the response variable Y given the covariates $\mathbf{U} = \mathbf{u}$ and $\mathbf{X} = \mathbf{x}$. For the aforementioned example, the log-link is used, $\mathbf{U}$ is the time covariate, and $\mathbf{X}$ denotes the levels of pollutants. The conditional distribution of the number of weekly hospital admissions given the covariates can be modeled reasonably as a Poisson distribution with the mean function given by (6.1.1).

In the least-squares setting, model (6.1.1) with the identity link was introduced by Cleveland, Grosse and Shyu (1992) and extended by Hastie and Tibshirani (1993) in various ways. Furthermore, a two-step estimation procedure was proposed by Fan and Zhang (2000) to deal with the situations where coefficient functions admit different degrees of smoothness. An advantage of model (6.1.1) is that by allowing the coefficients $\{a_j(\cdot)\}$ to depend on $\mathbf{U}$, the modeling bias can be reduced significantly and the "curse of dimensionality" is avoided.

Estimation of the coefficient functions in (6.1.1) is obtained by using local smoothing techniques. By localizing data around $\mathbf{u}$, model (6.1.1) is approximately a generalized linear model. One can find its local maximum likelihood estimate (MLE) using an iterative algorithm. Note that the local MLE for the varying-coefficient model is indeed solving the local likelihood equations. Thus, the proposed local likelihood method can be regarded as a special case of the general local estimation equation method proposed by Carroll, Ruppert and Welsh (1998). Hence, the bandwidth involved can be selected by the empirical bias method proposed in that paper. In order to obtain the estimated coefficient functions, one needs to solve hundreds of local maximum likelihood problems. The computation can be expensive, depending on the convergence criterion. Computational burden becomes even more severe when a cross-validation method is used to select a smoothing parameter. To reduce computational costs, we propose a one-step local MLE. The idea is not novel since it was first used by Bickel (1975) in the parametric setting, but implementations and insights are. It will be shown that computational costs can be reduced significantly and the resulting one-step estimator is demonstrated, both asymptotically and empirically, to be as efficient as the fully iterative MLE.

Associated with inferences on the varying-coefficient models are the standard errors of the estimated coefficient functions. Consistent estimates are derived. Simulation studies in Section 6.5 show that the estimated standard errors are very accurate for most applications. Another important issue arises regarding whether some of the coefficient functions in model (6.1.1) are actually varying, or whether some of covariates are statistically significant. A nonparametric maximum likelihood ratio test is proposed and its null distribution is estimated by using a conditional bootstrap method. Simulations in Section 6.5 show that the resulting testing

procedure performs well.

One of goals in this chapter is to estimate efficiently the coefficient functions $\{a_j(\cdot)\}$ in model (6.1.1) by using a nonparametric method. The proposed methods are directly applicable to situations in which one can not specify fully the conditional log-likelihood function $\ell(v,\, y)$, but can model the relationship between the mean and variance by $\text{var}(Y \mid \mathbf{U} = \mathbf{u},\, \mathbf{X} = \mathbf{x}) = \sigma^2 V\{m(\mathbf{u},\, \mathbf{x})\}$ for a known variance function $V(\cdot)$ and unknown σ. In this case, one needs only to replace the log-likelihood function $\ell(v,\, y)$ by the quasi-likelihood function $Q(\cdot,\, \cdot)$, defined by $\frac{\partial}{\partial \mu} Q(\mu, y) = \frac{y-\mu}{V(\mu)}$. It is assumed throughout this chapter that the conditional log-likelihood function $\ell(v,\, y)$ is known and linear in y for fixed v. This assumption is satisfied for the canonical exponential family, which is the focus of this chapter.

6.2 Estimation of Varying Coefficients

For simplicity, consider only the case that $\mathbf{u}$ is one-dimensional. Extension to multivariate $\mathbf{u}$ involves no fundamentally new ideas. However, implementations with $\mathbf{u}$ having more than two dimensions may encounter difficulties with the "curse of dimensionality," referred to the fact that a local neighborhood in higher dimensions is no longer local.

6.2.1 Local MLE

The local likelihood approach is used to estimate the varying-coefficients. Suppose that $a_j(\cdot)$ has a continuous second derivative. For each given point u_0, $a_j(u)$ is locally approximated by a linear function $a_j(u) \approx a_j + b_j\,(u - u_0)$ for u in a neighborhood of u_0. The logarithm of the local likelihood function based on a random sample $\{(U_i,\, \mathbf{X}_i,\, Y_i)\}_{i=1}^n$ is

$$\ell_n(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \ell \left[g^{-1} \left\{ \sum_{j=1}^{p} (a_j + b_j(U_i - u_0)) X_{ij} \right\},\, Y_i \right] K_h(U_i - u_0), \tag{6.2.1}$$

where $\mathbf{a} = (a_1,\, \ldots,\, a_p)^T$ and $\mathbf{b} = (b_1,\, \ldots,\, b_p)^T$. Note that a_j and b_j are dependent on u_0, and so is $\ell_n(\cdot,\, \cdot)$. Maximizing the local likelihood function $\ell_n(\mathbf{a},\, \mathbf{b})$ gives estimates $\hat{\mathbf{a}}(u_0)$ and $\hat{\mathbf{b}}(u_0)$. The components in $\hat{\mathbf{a}}(u_0)$ gives estimates of $a_1(u_0),\, \ldots,\, a_p(u_0)$. For simplicity of notation, we denote $\boldsymbol{\beta} = \boldsymbol{\beta}(u_0) = (a_1,\, \ldots,\, a_p,\, b_1,\, \ldots,\, b_p)^T$ and write the local likelihood function (6.2.1) as $\ell_n(\boldsymbol{\beta})$. Likewise, the local MLE is denoted by $\hat{\boldsymbol{\beta}}_{\text{MLE}} = \hat{\boldsymbol{\beta}}_{\text{MLE}}(u_0)$.

6.2.2 One-step local MLE

The local MLE can be costly to compute. This is particularly the case for the varying-coefficient models. In order to obtain the estimated functions $\{\hat{a}_j(\cdot)\}$, one needs to maximize the local likelihood (6.2.1) for usually hundreds of distinct values of u_0, with each maximization requiring an iterative algorithm. Moreover, the computational expense further increases with the number of covariates p. To ameliorate this expense, we propose to replace the iterative local MLE by the one-step Newton-Raphson estimator, which has been frequently used in parametric models (Bickel 1975; Lehmann 1983). Theorem 6.2 below shows that the one-step local MLE does not lose any statistical efficiency provided that the initial estimator is good enough.

Let $\ell_n'(\boldsymbol{\beta})$ and $\ell_n''(\boldsymbol{\beta})$ be the gradient and Hessian matrix of the local log-likelihood $\ell_n(\boldsymbol{\beta})$. Given an initial estimator $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_0(u_0) = \left(\hat{\mathbf{a}}(u_0)^T,\, \hat{\mathbf{b}}(u_0)^T\right)^T$, the one-step of the Newton-Raphson algorithm produces the updated estimator,

$$\hat{\boldsymbol{\beta}}_{\mathrm{OS}} = \hat{\boldsymbol{\beta}}_0 - \left\{\ell_n''\left(\hat{\boldsymbol{\beta}}_0\right)\right\}^{-1} \ell_n'\left(\hat{\boldsymbol{\beta}}_0\right), \tag{6.2.2}$$

thus featuring the computational expediency of least-squares local polynomial fitting. An approach to constructing a good initial estimator is stated in Section 6.2.4.

Note that $\ell_n''(\hat{\boldsymbol{\beta}}_0)$ can be nearly singular for certain u_0, due to possible data sparsity in some regions, or when the bandwidth is too small. Seifert and Gasser (1996) and Fan and Chen (1999) explored the use of ridge regression as an approach to handling such problems in the univariate setting. Their ideas are modified for the current setting in this chapter.

6.2.3 Standard errors

Since the local likelihood (6.2.1) is a weighted likelihood function of a parametric generalized linear model, the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ can be estimated using conventional techniques. Let $q_j(s,\, y) = (\partial^j/\partial s^j)\, \ell\,\{g^{-1}(s),\, y\}$ and

$$\hat{\Gamma}(u_0) = -\,\frac{1}{n}\sum_{i=1}^{n} q_2\left[\sum_{j=1}^{p}\{\hat{a}_j(u_0)\, X_{ij} + \hat{b}_j(u_0)(U_i - u_0)\},\; Y_i\right] K_h(U_i - u_0) \begin{pmatrix} \mathbf{X}_i \\ \mathbf{X}_i\,(U_i - u_0)/h \end{pmatrix}^{\otimes 2}, \tag{6.2.3}$$

where $A^{\otimes 2}$ denotes $A\,A^T$ for a matrix or vector A. Then, the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$ can be estimated as

$$\hat{\Sigma}^*(u_0) = \hat{\Gamma}(u_0)^{-1}\,\hat{\Lambda}(u_0)\,\hat{\Gamma}(u_0)^{-1}, \tag{6.2.4}$$

where

$$\hat{\Lambda}(u_0) = \frac{h}{n} \sum_{i=1}^{n} q_1^2 \left[\sum_{j=1}^{p} \{\hat{a}_j(u_0)\, X_{ij} + \hat{b}_j(u_0)(U_i - u_0)\},\ Y_i \right] K_h^2(U_i - u_0) \begin{pmatrix} \mathbf{X}_i \\ \mathbf{X}_i\,(U_i - u_0)/h \end{pmatrix}^{\otimes 2}.$$

In the implementation in Section 6.5, a ridge regression technique is employed and hence the matrix $\hat{\Gamma}(u_0)$ in (6.2.4) is slightly modified to reflect this change.

The explicit formula for the asymptotic covariance matrix in (6.3.7) below provides an alternative estimate of the asymptotic covariance matrix of $\mathbf{a}(u_0)$ (not full vector $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$) $\Sigma(u_0)$. Therefore, a direct estimate of $\Sigma(u_0)$ is $\widetilde{\Sigma}(u_0) = \nu_0\, \hat{\Gamma}_S(u_0)^{-1}$, where $\hat{\Gamma}_S(u_0)$ is the $p \times p$ upper corner submatrix of $\hat{\Gamma}(u_0)$ given by (6.2.3).

6.2.4 Implementation of one-step local MLE

Suppose that we wish to evaluate the functions $\hat{\mathbf{a}}(\cdot)$ at grid points $u_j,\ j = 1, \ldots, n_{\mathrm{grid}}$. Our idea of finding initial estimators is as follows. Take a point u_{i_0}, usually the center of the grid points. Compute the local MLE $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_{i_0})$. Use this estimate as the initial estimate for the point u_{i_0+1} and apply (6.2.2) to obtain $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0+1})$. Now, use $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0+1})$ as the initial estimate at the point u_{i_0+2} and apply (6.2.2) to obtain $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0+2})$ and so on. Likewise, we can compute $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0-1})$, $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{i_0-2})$, etc. In this way, we obtain our estimates at all grid points.

There are a couple of possible variations to the above technique. The first one is to calculate a fresh local MLE as a new initial value after iterating along the grid points for a while. For example, if we wish to evaluate the functions at 200 grid points and are willing to compute the local maximum likelihood at five distinct points. A sensible placement of these points is u_{20}, u_{60}, u_{100}, u_{140} and u_{180}. Use for example $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_{60})$ along with the idea in the last paragraph to compute $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_i)$ for $i = 40, \ldots, 79$. In our implementation, this modified technique is used.

Another useful modification is to use a two-step method. We use the scenarios given in the last paragraph as an illustration. After obtaining $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_{60})$, say, we apply (6.2.2) to obtain $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{61})$. Regarding $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}(u_{61})$ as an initial value, we use (6.2.2) to obtain a "two-step" estimator $\hat{\boldsymbol{\beta}}_{\mathrm{TS}}(u_{61})$. Now, use $\hat{\boldsymbol{\beta}}_{\mathrm{TS}}(u_{61})$ as an initial value for the grid point u_{62} and iterate (6.2.2) twice to obtain $\hat{\boldsymbol{\beta}}_{\mathrm{TS}}(u_{62})$ and so on. This implementation requires approximately twice as much effort to compute the estimates as the one-step method. However, our empirical

studies show that there are no significant differences between the two procedures. See Section 6.5 for details.

The theoretical basis for the above "one-step" and the "two-step" procedures is as follows. When the grid points are sufficiently fine, $\hat{\boldsymbol{\beta}}_{\rm MLE}(u_{i_0})$ will be very close to $\hat{\boldsymbol{\beta}}_{\rm MLE}(u_{i_0+1})$. Indeed, when the grid span is of the order $O\left\{h_n^2 + (n\,h_n)^{-1/2}\right\}$ which usually is true for most applications, $\hat{\boldsymbol{\beta}}_{\rm MLE}(u_{i_0})$ satisfies the condition given in Theorem 6.2. Therefore, $\hat{\boldsymbol{\beta}}_{\rm OS}(u_{i_0+1})$ is as efficient as the fully-iterative local MLE at the point u_{i_0+1}. Using the same reasoning, $\hat{\boldsymbol{\beta}}_{\rm OS}(u_{i_0+2})$ is as efficient as the local MLE at the point $u = u_{i_0+2}$ and so on. The same arguments are still applicable for the two-step estimator. A refresh start is needed because of stochastic error accumulation as iterations along grid points march on.

Based on the above theoretical considerations, we suggest a very simple rule of thumb for choosing the number of grid points: $n_{\rm grid} = \max\{200, {\rm IQR}^2/h^2\}$, where IQR is the interquantile range of $U_1, \cdots, U_n$. In such a way, approximation errors between estimates at two consecutive grid points are of order $O(h^2)$, satisfying the critical condition (6.3.20).

6.3 Main Results

Define $\mu_k = \int u^k\, K(u)\, d\, u$ and $\nu_k = \int u^k\, K^2(u)\, du$. Let $\mathbf{H} = {\rm diag}(1,\, h) \otimes \mathbf{I}_p$ with $\otimes$ denoting the Kronecker product. Let $f_{\rm U}(\cdot)$ denote the marginal density of U,

$$\Gamma(u) = E\left\{\rho(U,\, \mathbf{X})\, \mathbf{X}\, \mathbf{X}^T \,|\, U = u\right\}, \tag{6.3.1}$$

and

$$\rho(u,\, \mathbf{x}) = [g_1\{m(u,\, \mathbf{x})\}]^2\ {\rm var}\{Y|U = u,\, \mathbf{X} = \mathbf{x}\} \tag{6.3.2}$$

with $g_1(s) = g_0'(s)/g'(s)$ and $g_0(\cdot)$ being the canonical link. Note that $\rho(u,\, \mathbf{x}) = V\{m(u,\, \mathbf{x})\}$ for the canonical exponential family with the canonical link function. The asymptotic properties of $\hat{\boldsymbol{\beta}}_{\rm MLE}$ and $\hat{\boldsymbol{\beta}}_{\rm OS}$ are described in the following theorems. Carroll, Ruppert and Welsh (1998) studied local estimating equations. Local likelihood approaches may be regarded as a special case of local estimating equations because the local maximum likelihood estimates are the solution of local likelihood equations. They obtained expressions of asymptotic bias and variance for the solution of local estimating equations. Theorems 6.1 and 6.2 present the asymptotic results in terms of asymptotic normality of the local maximum likelihood estimates.

Let $q_j(s,\, y) = (\partial^j/\partial s^j)\, \ell\, \{g^{-1}(s),\, y\}$. Note that $q_k(s,\, y)$ is linear in y for fixed s such that

$$q_1[g\{m(u,\, \mathbf{x})\},\, m(u,\, \mathbf{x})] = 0 \quad \text{and} \quad q_2[g\{m(u,\, \mathbf{x})\},\, m(u,\, \mathbf{x})] = -\rho(u,\, \mathbf{x}),\,. \tag{6.3.3}$$

We first state some regularity conditions.

Conditions:

(1) The function $q_2(s,\, y) < 0$ for $s \in \Re$ and y in the range of the response variable.

(2) The functions $f_{\mathrm{U}}(u)$, $\Gamma(u)$, $V(m(u,\, \mathbf{x}))$, $V'(m(u,\, \mathbf{x}))$ and $g'''(m(u,\, \mathbf{x}))$ are continuous at the point $u = u_0$. Further, assume that $f_{\mathrm{U}}(u_0) > 0$ and $\Gamma(u_0) > 0$.

(3) $K(\cdot)$ has a bounded support.

(4) $a_j''(\cdot)$ is continuous in a neighborhood of u_0 for $j = 1,\, \ldots,\, p$.

(5) $E\,\{|\mathbf{X}|^3 \,|\, U = u\}$ is continuous at the point $u = u_0$.

(6) $E(Y^4 \,|\, U = u,\, \mathbf{X} = \mathbf{x})$ is bounded in a neighborhood of $u = u_0$.

Condition (1) guarantees that the local likelihood function (6.2.1) is concave. It is satisfied for the canonical exponential family with a canonical link. Note that Condition (2) implies that $q_1(\cdot, \cdot)$, $q_2(\cdot, \cdot)$, $q_3(\cdot, \cdot)$, $\rho'(\cdot, \cdot)$ and $m'(\cdot, \cdot)$ are continuous.

Theorem 6.1 *Suppose that Conditions (1) – (6) hold and that $h = h_n \to 0$ and $n\,h \to \infty$ as $n \to \infty$. Then*

$$\begin{aligned} &\sqrt{n\,h}\left[\mathbf{H}\left\{\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}(u_0) - \boldsymbol{\beta}(u_0)\right\} - \frac{h^2}{2(\mu_2 - \mu_1^2)}\begin{pmatrix}(\mu_2^2 - \mu_1\,\mu_3)\,\mathbf{a}''(u_0)\\ (\mu_3 - \mu_1\,\mu_2)\,\mathbf{a}''(u_0)\end{pmatrix} + o_p(h^2)\right] \\ &\quad \xrightarrow{\mathcal{D}} \quad N\left(0,\ \Delta^{-1}\,\Lambda\,\Delta^{-1}\right) \end{aligned} \tag{6.3.4}$$

with $\Gamma(u_0)$ given by (6.3.1),

$$\Delta = f_{\mathrm{U}}(u_0)\begin{pmatrix}1 & \mu_1\\ \mu_1 & \mu_2\end{pmatrix} \otimes \Gamma(u_0) \quad \text{and} \quad \Lambda = f_{\mathrm{U}}(u_0)\begin{pmatrix}\nu_0 & \nu_1\\ \nu_1 & \nu_2\end{pmatrix} \otimes \Gamma(u_0). \tag{6.3.5}$$

Furthermore, if $K(\cdot)$ is symmetric,

$$\sqrt{n\,h}\ \left[\hat{\mathbf{a}}_{\mathrm{MLE}}(u_0) - \mathbf{a}(u_0) - \frac{h^2\,\mu_2}{2}\,\mathbf{a}''(u_0) + o_p(h^2)\right] \quad \xrightarrow{\mathcal{D}} \quad N\,(0,\ \Sigma(u_0))\,, \tag{6.3.6}$$

where

$$\Sigma(u_0) = \nu_0\,\Gamma^{-1}(u_0)/f_{\mathrm{U}}(u_0). \tag{6.3.7}$$

Proof: Recall that $\hat{\boldsymbol{\beta}}_{\rm MLE}$ maximizes (6.2.1). Let $\overline{\eta}(u_0,\, u,\, \mathbf{x}) = \sum_{j=1}^{p}\{a_j(u_0) + a_j'(u_0)(u - u_0)\}\, x_j$, and

$$\boldsymbol{\beta}^* = \gamma_n^{-1}\, \Big(\beta_1 - a_1(u_0),\, \ldots,\, \beta_p - a_p(u_0),\, h(\beta_{p+1} - a_1'(u_0)),\, \ldots,\, h(\beta_{2p} - a_p'(u_0))\Big)^T,$$

where $\gamma_n = (n\,h)^{-1/2}$. It can easily be seen that $\sum_{j=1}^{p}\{a_j + b_j(U_i - u_0)\}\, X_{ij} = \overline{\eta}(u_0,\, U_i,\, \mathbf{X}_i) + \gamma_n\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i$, where $\mathbf{Z}_i = \Big(\mathbf{X}_i^T,\, ((U_i - u_0)/h)\, \mathbf{X}_i^T\Big)^T$. Then, the local likelihood function $\ell_n(\boldsymbol{\beta})$ defined in (6.2.1) becomes

$$\ell_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left[g^{-1}\left\{\overline{\eta}(u_0,\, U_i,\, \mathbf{X}_i) + \gamma_n\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right\},\, Y_i\right] K_h(U_i - u_0),$$

which is a function of $\boldsymbol{\beta}^*$, denoted by $\ell_n(\boldsymbol{\beta}^*)$. Let

$$\hat{\boldsymbol{\beta}}^{\,*} = \gamma_n^{-1}\, \Big(\hat{\beta}_1 - a_1(u_0),\, \ldots,\, \hat{\beta}_p - a_p(u_0),\, h\left(\hat{\beta}_{p+1} - a_1'(u_0)\right),\, \ldots,\, h\left(\hat{\beta}_{2p} - a_p'(u_0)\right)\Big)^T.$$

Then $\hat{\boldsymbol{\beta}}^{\,*}$ maximizes $\ell_n(\boldsymbol{\beta}^*)$ since $\hat{\boldsymbol{\beta}}$ maximizes (6.2.1). Equivalently, $\hat{\boldsymbol{\beta}}^{\,*}$ maximizes the following normalized function

$$\ell_n^*(\boldsymbol{\beta}^*) = \sum_{i=1}^{n}\left[\ell\left\{g^{-1}\left(\overline{\eta}_i(u_0) + \gamma_n\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right),\, Y_i\right\} - \ell\left\{g^{-1}\left(\overline{\eta}_i(u_0)\right),\, Y_i\right\}\right] K\left\{(U_i - u_0)/h\right\},$$

where $\overline{\eta}_i(u_0) = \overline{\eta}(u_0,\, U_i,\, \mathbf{X}_i)$.

Note that Condition (1) implies that $\ell_n^*(\cdot)$ is concave in $\boldsymbol{\beta}^*$. Using the Taylor expansion of $\ell\,\{g^{-1}(\cdot),\, y\}$, we have

$$\ell_n^*(\boldsymbol{\beta}^*) = W_n^T\, \boldsymbol{\beta}^* + \frac{1}{2}\, \boldsymbol{\beta}^{*T}\, \Delta_n\, \boldsymbol{\beta}^* + \frac{\gamma_n^3}{6}\sum_{i=1}^{n} q_3\,\{\eta_i,\, Y_i\}\, \left(\boldsymbol{\beta}^{*T}\, \mathbf{Z}_i\right)^3\, K\left\{(U_i - u_0)/h\right\}, \qquad (6.3.8)$$

where

$$W_n = \gamma_n\, \sum_{i=1}^{n} q_1\,\{\overline{\eta}_i(u_0),\, Y_i\}\, \mathbf{Z}_i\, K\left\{(U_i - u_0)/h\right\}, \qquad (6.3.9)$$

$$\Delta_n = \frac{\gamma_n^2}{2}\sum_{i=1}^{n} q_2\,\{\overline{\eta}_i(u_0),\, Y_i\}\, \mathbf{Z}_i\, \mathbf{Z}_i^T\, K\left\{(U_i - u_0)/h\right\},$$

and η_i is between $\overline{\eta}_i(u_0)$ and $\overline{\eta}_i(u_0) + \gamma_n\, \boldsymbol{\beta}^{*T}\, \mathbf{Z}_i$. Note that

$$(\Delta_n)_{ij} = (E\Delta_n)_{ij} + O_p\left[\{\mathrm{var}(\Delta_n)_{ij}\}^{1/2}\right].$$

Now the mean in the above expression equals

$$E(\Delta_n) = h^{-1}\, E\left[q_2\,\{\overline{\eta}(u_0,\, U,\, \mathbf{X}),\, m(U,\, \mathbf{X})\}\, K\left\{(U - u_0)/h\right\}\, \mathbf{Z}\, \mathbf{Z}^T\right].$$

By a Taylor series expansion of $\eta(u,\, \mathbf{x})$ with respect to u around $|u - u_0| < h$ and the first result in (6.3.3), we have

$$\eta(u,\, \mathbf{x}) = \overline{\eta}(u_0,\, u,\, \mathbf{x}) + \frac{h^2\,(u-u_0)^2}{2}\,\eta''_u(u_0,\, \mathbf{x}) + o(h^2),$$

where $\eta''_u(u,\, \mathbf{x}) = (\partial^2/\partial u^2)\eta(u,\, \mathbf{x}) = \sum_{j=1}^{p}\, a''_j(u)\, x_j$, which implies that

$$q_1\{\overline{\eta}(u_0,\, u,\, \mathbf{x}),\ m(u,\, \mathbf{x})\} = \rho(u,\, \mathbf{x})\,\frac{h^2\,(u-u_0)^2}{2}\,\eta''_u(u_0,\, \mathbf{x}) + o(h^2), \tag{6.3.10}$$

and

$$q_2\{\overline{\eta}(u_0,\, u,\, \mathbf{x}),\ m(u,\, \mathbf{x})\} = -\rho(u,\, \mathbf{x}) + o(1). \tag{6.3.11}$$

Then, using the second equality of (6.3.3) and (6.3.11), we obtain

$$E(\Delta_n)\ \rightarrow\ -\,f_{\mathrm{U}}(u_0)\,\begin{pmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix} \otimes\, \Gamma(u_0) = -\,\Delta, \tag{6.3.12}$$

where $\Gamma(u_0)$ is given in (6.3.1) and Δ is defined in (6.3.5). Similar arguments show that $\mathrm{var}\{(\Delta_n)_{ij}\} = O\,\{(n\,h)^{-1}\}$. Therefore,

$$\Delta_n = -\Delta + o_p(1). \tag{6.3.13}$$

Since $K(\cdot)$ is bounded, $q_3(\cdot,\, \cdot)$ is linear in Y_1 and $E(|Y_1|\,|\,U_1,\, \mathbf{X}_1) < \infty$, the expected value of the absolute value of the last term in (6.3.8) is bounded by

$$O\left(n\,\gamma_n^3\,E\Big|q_3(\eta_1,\, Y_1)\,\mathbf{X}_1^3\,K\,\{(U_1-u_0)/h\}\,\Big|\right) = O(\gamma_n) \tag{6.3.14}$$

by Condition (5). Therefore, the last term in (6.3.8) is of order $O_p(\gamma_n)$. This, in conjunction with (6.3.8), (6.3.12) and (6.3.13), implies that

$$\ell_n^*(\boldsymbol{\beta}^*) = W_n^T\,\boldsymbol{\beta}^* - \frac{1}{2}\,\boldsymbol{\beta}^{*T}\,\Delta\,\boldsymbol{\beta}^* + o_p(1).$$

An application of the quadratic approximation lemma (see, for example, Fan and Gijbel 1996, p.210) leads to

$$\hat{\boldsymbol{\beta}}^{\,*} = \Delta^{-1}\,W_n + o_p(1), \tag{6.3.15}$$

if W_n is a sequence of stochastically bounded random vectors. The asymptotic normality of $\hat{\boldsymbol{\beta}}^{\,*}$ follows from that of W_n. Hence, it remains to establish the asymptotic normality of W_n.

Note that the random vector W_n is a sum of i.i.d. random vectors. In order to establish its asymptotic normality, it suffices to compute the mean and covariance matrix of W_n and check

the Lyapounov condition. To this end, by (6.3.10), we have

$$\begin{aligned} E(W_n) &= n\,\gamma_n\,E\left[q_1\left\{\overline{\eta}(u_0,\,U,\,\mathbf{X}),\,m(U,\,\mathbf{X})\right\}\,\mathbf{Z}\,K\left\{(U-u_0)/h\right\}\right] \\ &= \frac{h^2\,f_{\rm U}(u_0)}{2\,\gamma_n}\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}\otimes\Gamma(u_0)\,\mathbf{a}''(u_0)\,\left\{1+o(1)\right\}. \end{aligned} \tag{6.3.16}$$

Similarly, by (6.3.16) and the definition of $q_1(\cdot,\,\cdot)$, one has

$$\begin{aligned} \mathrm{var}(W_n) &= h^{-1}\,E\left[q_1^2\left\{\overline{\eta}(u_0,\,U,\,\mathbf{X}),\,Y\right)\right\}\,\mathbf{Z}\,\mathbf{Z}^T\,K^2\left\{(U-u_0)/h\right\}\Big] \\ &= f_{\rm U}(u_0)\begin{pmatrix}\nu_0 & \nu_1\\ \nu_1 & \nu_2\end{pmatrix}\otimes\Gamma(u_0)\,\left\{1+o(1)\right\}=\Lambda+o(1), \end{aligned} \tag{6.3.17}$$

where Λ is defined in (6.3.5). By the Cramér-Wold device, in order to derive the asymptotic normality of W_n, it suffices to show that for any unit vector $\mathbf{d}\in\Re^{2p}$,

$$\left\{\mathbf{d}^T\,\mathrm{var}(W_n)\,\mathbf{d}\right\}^{-1/2}\,\left\{\mathbf{d}^T\,W_n-\mathbf{d}^T\,E(W_n)\right\}\quad\stackrel{\mathcal{D}}{\longrightarrow}\quad N(0,\,1). \tag{6.3.18}$$

This, conjunction with (6.3.15), (6.3.16), and (6.3.17), implies that

$$\hat{\boldsymbol{\beta}}^*-\frac{(n\,h^5)^{1/2}}{2}\,\Delta^{-1}\,f_{\rm U}(u_0)\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}\otimes\Gamma(u_0)\,\mathbf{a}''(u_0)\,\left\{1+o(1)\right\}\quad\stackrel{\mathcal{D}}{\longrightarrow}\quad N\left(0,\,\Delta^{-1}\,\Lambda\,\Delta^{-1}\right). \tag{6.3.19}$$

Therefore, the assertion in (6.3.4) holds true. To prove (6.3.18), we need only to check Lyapounov's condition for that sequence. To do so, let $\xi_i=q_1\left\{\overline{\eta}_i(u_0),\,Y_i\right\}\,\mathbf{d}^T\,\mathbf{Z}_i\,K\left\{(U_i-u_0)/h\right\}$. Then, $\mathbf{d}^T\,W_n=\gamma_n\,\sum_{i=1}^n\xi_i$. It suffices to show that $n\,\gamma_n^3\,E|\xi_1|^3\to 0$ as $n\to\infty$. Similar to (6.3.14), one can show that $n\,\gamma_n^3\,E|\xi_1|^3=O(\gamma)\to 0$. If $K(\cdot)$ is symmetric, then $\mu_1=0$, so that (6.3.6) holds true. This completes the proof of the theorem. □

Theorem 6.2 *Under the assumptions in Theorem 1, then $\hat{\boldsymbol{\beta}}_{\rm OS}$ has the same asymptotic distribution as $\hat{\boldsymbol{\beta}}_{\rm MLE}$, provided that the initial estimator $\hat{\boldsymbol{\beta}}_0$ satisfies*

$$\mathbf{H}\,\left(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}\right)=O_p\left\{h^2+(n\,h)^{-1/2}\right\}. \tag{6.3.20}$$

Proof: Recall that $\ell_n(\boldsymbol{\beta})=\frac{1}{n}\,\sum_{i=1}^n\ell\left\{g^{-1}\left(\sum_{j=1}^p(a_j+b_j\,(U_i-u_0))\,X_{ij}\right),\,Y_i\right\}K_h(U_i-u_0)$. For any $\widetilde{\boldsymbol{\beta}}$ satisfying $\mathbf{H}\left(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)=O_p\left(h^2+(n\,h)^{-1/2}\right)$, one can easily show that

$$\begin{aligned} \mathbf{H}^{-1}\,\ell_n''\left(\widetilde{\boldsymbol{\beta}}\right)\,\mathbf{H}^{-1} &= \mathbf{H}^{-1}\,\ell_n''(\boldsymbol{\beta})\,\mathbf{H}^{-1}+o_p(1) \\ &= \frac{1}{n}\,\sum_{i=1}^n q_2\left\{\widetilde{\mathbf{Z}}_i^T\boldsymbol{\beta},\,Y_i\right\}\,\mathbf{H}^{-1}\,\widetilde{\mathbf{Z}}_i\,\widetilde{\mathbf{Z}}_i^T\,\mathbf{H}^{-1}\,K_h(U_i-u_0)+o_p(1), \end{aligned} \tag{6.3.21}$$

where $\widetilde{\mathbf{Z}}_i = \left(\mathbf{X}_i^T, (U_i - u_0)\mathbf{X}_i^T\right)^T$. By computing the mean and variance of $\mathbf{H}^{-1}\,\ell_n''(\boldsymbol{\beta})\,\mathbf{H}^{-1}$, we obtain

$$\begin{aligned}
\mathbf{H}^{-1}\,\ell_n''\left(\widetilde{\boldsymbol{\beta}}\right)\,\mathbf{H}^{-1} &= E\left[q_2\left\{\widetilde{\mathbf{Z}}^T\boldsymbol{\beta},\ Y\right\}\begin{pmatrix}1\\(U-u_0)/h\end{pmatrix}^{\otimes 2}\otimes \mathbf{X}\,\mathbf{X}^T\,K_h(U-u_0)\right] + o_p(1)\\
&= E\left[q_2\left\{\widetilde{\mathbf{Z}}^T\boldsymbol{\beta},\ m(U,\,\mathbf{X})\right\}\begin{pmatrix}1\\(U-u_0)/h\end{pmatrix}^{\otimes 2}\otimes \mathbf{X}\,\mathbf{X}^T\,K_h(U-u_0)\right] + o_p(1)\\
&= -\Delta + o_p(1), \qquad (6.3.22)
\end{aligned}$$

where Δ is defined in (6.3.5). Recall that $\hat{\boldsymbol{\beta}}_{\mathrm{OS}} = \hat{\boldsymbol{\beta}}_0 - \left\{\ell_n''\left(\hat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\,\ell_n'\left(\hat{\boldsymbol{\beta}}_0\right)$ (see (6.2.2)). By the Taylor expansion, we have

$$\ell_n'\left(\hat{\boldsymbol{\beta}}_0\right) = \ell_n'(\boldsymbol{\beta}) + \ell_n''\left(\widetilde{\boldsymbol{\beta}}^*\right)\,\left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right),$$

where $\widetilde{\boldsymbol{\beta}}^*$ lies between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_0$ and hence satisfies $\mathbf{H}\,\left(\widetilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta}\right) = O_p\left(h^2 + (n\,h)^{-1/2}\right)$. Then, some algebraic computations show that

$$\begin{aligned}
\mathbf{H}\,\left(\hat{\boldsymbol{\beta}}_{\mathrm{OS}} - \boldsymbol{\beta}\right) &= \mathbf{H}\,\left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right) - \mathbf{H}\,\left\{\ell_n''\left(\hat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\,\mathbf{H}\,\mathbf{H}^{-1}\,\ell_n'\left(\hat{\boldsymbol{\beta}}_0\right)\\
&= \left[\mathbf{I} - \mathbf{H}\,\left\{\ell_n''\left(\hat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\,\mathbf{H}\,\mathbf{H}^{-1}\,\ell_n''\left(\widetilde{\boldsymbol{\beta}}^*\right)\,\mathbf{H}^{-1}\right]\,\mathbf{H}\,\left(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}\right)\\
&\quad - \mathbf{H}\,\left\{\ell_n''\left(\hat{\boldsymbol{\beta}}_0\right)\right\}^{-1}\,\mathbf{H}\,\mathbf{H}^{-1}\,\ell_n'(\boldsymbol{\beta}). \qquad (6.3.23)
\end{aligned}$$

Therefore, by (6.3.22) and (6.3.23), we have

$$\mathbf{H}\,\left(\hat{\boldsymbol{\beta}}_{\mathrm{OS}} - \boldsymbol{\beta}\right) = \Delta^{-1}\,\mathbf{H}^{-1}\,\ell_n'(\boldsymbol{\beta})\,\{1 + o_p(1)\} + o_p\left(h^2 + (n\,h)^{-1/2}\right),$$

which, in conjunction with (6.3.9), (6.3.15), (6.3.18) and (6.3.19), implies that

$$\sqrt{n\,h}\,\mathbf{H}\,\left(\hat{\boldsymbol{\beta}}_{\mathrm{OS}} - \boldsymbol{\beta}\right) = \Delta^{-1}\,W_n + o_p(1) = \hat{\boldsymbol{\beta}}^* + o_p(1). \qquad (6.3.24)$$

Therefore, $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}$ has the same asymptotic distribution as $\hat{\boldsymbol{\beta}}_{\mathrm{MLE}}$. □

As a consequence, the fully iterative MLE and the one-step estimate share the same asymptotic properties provided that (6.3.20) is fulfilled, which provide the theoretical basis for the use of the one-step approach in practice. The asymptotic mean squared error (MSE) of two estimators $\hat{a}_{j,\mathrm{MLE}}(u_0)$ and $\hat{a}_{j,\mathrm{OS}}(u_0)$ is

$$\mathrm{MSE} = \frac{h^4}{4}\,\mu_2^2\,\left\{a_j''(u_0)\right\}^2 + \frac{\nu_0\,\sigma_{jj}^2(u_0)}{n\,h\,f_U(u_0)},$$

when $K(\cdot)$ is symmetric, where $\sigma_{jj}^2(u_0)$ is the j-th diagonal element of $\Gamma^{-1}(u_0)$. Then, the MSE is of order $n^{-4/5}$ if the optimal bandwidth is used.

6.4 Hypothesis Testing for Varying Coefficients

When fitting a varying-coefficient model, one naturally asks whether the coefficient functions are actually varying or whether any particular covariate is significant in the model. For simplicity of description, we only consider the simple hypothesis testing problem

$$H_0 : a_1(u) \equiv a_1, \cdots, a_p(u) \equiv a_p, \tag{6.4.1}$$

though the technique also applies to other testing problems. A useful procedure is based on the nonparametric likelihood ratio test statistic

$$T = 2\{\ell(H_1) - \ell(H_0)\}, \tag{6.4.2}$$

where $\ell(H_0)$ and $\ell(H_1)$ are respectively the log-likelihood functions computed under the null and alternative hypotheses.

For parametric models, the likelihood ratio statistic follows asymptotically a χ^2-distribution with degrees of freedom $f - r$, where r and f are the number of parameters under the null and alternative hypotheses. For the nonparametric alternative, the effective number of parameters f tends to infinity. Thus, the test statistic will be asymptotically normal, *independent* of the values $a_1, \cdots, a_p$. For rigorous justification, see the paper by Fan, Zhang and Zhang (1999) who considered sieve likelihood ratio tests in a general setting and demonstrated that the Wilks' type of phenomenon holds for a large variety of nonparametric problems. This in turn suggests that we can use the following *conditional bootstrap* to construct the null distribution of T. Let $\{\hat{a}_j\}$ be the MLE under the null hypothesis. Given the covariates $(U_i, \mathbf{X}_i)$, generate a bootstrap sample Y_i^* from the given distribution of Y with the estimated linear predictor $\hat{\eta}(U_i, \mathbf{X}_i) = \sum_{j=1}^p \hat{a}_j X_{ij}$ and compute the test statistic T^* in (6.4.2). Use the distribution of T^* as an approximation to the distribution of T. This method is valid since the asymptotic null distribution does not depend on the values $\{a_j\}$ (Fan, Zhang and Zhang, 1999).

Note that the above conditional bootstrap method applies readily to the Poisson and Bernoulli distributions, since in these cases the distribution of Y does not involve any dispersion parameters. It is really a simulation approximation to the conditional distribution of T given observed covariates under the particular null hypothesis: $H_0 : a_j(u) = \hat{a}_j$ $(j = 1, \cdots, p)$. As pointed out above, this approximation is valid under both H_0 and H_1 as the null distribution does not asymptotically depend on the values of $\{a_j\}$. In the case where model (6.1.1) involves a dispersion parameter (e.g., the Gaussian model), the dispersion parameter should

be estimated based on the residuals from the *alternative* hypothesis. This is again due to the Wilks type of results demonstrated by Fan, Zhang and Zhang (1999).

For testing the hypothesis such as $a_p(\cdot) = 0$, the above conditional bootstrap idea continues to apply. In this case, the data should be generated from the mean function $g\{m(\mathbf{u}, \mathbf{x})\} = \sum_{j=1}^{p-1} \hat{a}_j(\mathbf{u})\, x_j$, where $\hat{a}_j(\cdot)$ is an estimate under the alternative hypothesis.

6.5 Simulations and Applications

In this section, we first discuss how to implement the one-step procedure for the Bernoulli and Poisson models. We then illustrate the performance of the proposed one-step method and compare it with the two-step estimator and the fully-iterative local MLE. The performance of estimator $\hat{\mathbf{a}}(\cdot)$ is assessed via the square-Root of Average Square Errors (RASE)

$$\text{RASE}^2 = n_{\text{grid}}^{-1} \sum_{j=1}^{p} \sum_{k=1}^{n_{\text{grid}}} \{\hat{a}_j(u_k) - a_j(u_k)\}^2, \tag{6.5.1}$$

where $\{u_k,\, k = 1, \ldots, n_{grid}\}$ are the grid points at which the functions $\{a_j(\cdot)\}$ are estimated.

In the following two simulated examples, the covariates X_1 and X_2 are standard normal random variables with correlation coefficient $2^{-1/2}$ and U is uniformly distributed over $[0, 1]$, independent of $(X_1,\, X_2)$. Three bandwidths will be employed to represent widely varying degrees of smoothness. Over this range of bandwidths, we compare the performances among the one-step, the two-step and the fully iterative local MLE methods. The Epanechnikov kernel $K(u) = 0.75(1 - u^2)_+$ and $n_{\text{grid}} = 200$ are used.

6.5.1 Logistic Regression

For a Bernoulli distribution, the one-step estimator is given by

$$\hat{\boldsymbol{\beta}}_{\text{OS}} = \hat{\boldsymbol{\beta}}_0 + \begin{pmatrix} \mathbf{H}_{n,0}, & \mathbf{H}_{n,1} \\ \mathbf{H}_{n,1}, & \mathbf{H}_{n,2} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{v}_{n,0} \\ \mathbf{v}_{n,1} \end{pmatrix}, \tag{6.5.2}$$

where $\mathbf{H}_{n,j} = \sum_{i=1}^{n} K_h(U_i - u_0)\hat{p}_{i0}(1-\hat{p}_{i0})(U_i - u_0)^j \mathbf{X}_i \mathbf{X}_i^T$, $j = 0,\, 1,\, 2$, $\hat{p}_{i0}$ satisfies $\text{logit}\,(\hat{p}_{i0}) = \sum_{j=1}^{p} \left\{\hat{a}_{j,0} + \hat{b}_{j,0}(U_i - u_0)\right\} X_{ij}$, and $\mathbf{v}_{n,j} = \sum_{i=1}^{n} K_h(U_i - u_0)(Y_i - \hat{p}_{i0})(U_i - u_0)^j \mathbf{X}_i$, $j = 0,\, 1$. The two-step estimator $\hat{\boldsymbol{\beta}}_{\text{TS}}$ is obtained by iterating the equation (6.5.2) twice and the local MLE is obtained by iterating equation (6.5.2) until convergence.

In practice, the matrix in (6.5.2) can be singular or nearly singular when the local data are sparse. To attenuate this difficulty, one may follow the idea of ridge regression (Seifert and Gasser 1996; Fan and Chen 1999). Then an issue arises on how to choose the ridge parameters. Note that the k-th diagonal element of $\mathbf{H}_{n,j}$ ($j = 0$ and 2) is approximately of order

$$E\left(X_k^2 \,|\, U = u_0\right) \hat{p}_0(1-\hat{p}_0)\, h^{j-1} \int u^j K(u)\, du\; N \qquad \text{with} \qquad \hat{p}_0 = \frac{\exp(\hat{\mathbf{a}}_0^T \overline{\mathbf{X}})}{1+\exp(\hat{\mathbf{a}}_0^T \overline{\mathbf{X}})}, \quad (6.5.3)$$

where $N = n\,h\,f_{\mathrm{U}}(u_0)$ and $\overline{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$. The parameter N can be intuitively understood as the effective number of local data points. This motivates us to use the ridge parameter

$$r_{j,k} = \left(\frac{1}{n}\sum_{i=1}^n X_{ik}^2\right) \hat{p}_0(1-\hat{p}_0) h^{j-1} \int u^j K(u)\, du$$

for the k-th diagonal element of $\mathbf{H}_{n,j}$. Using such a ridge parameter will not alter the asymptotic behavior and will prevent the matrix from becoming nearly singular when N is small. However, it affects and indeed ameliorates the finite-sample properties of estimators for small-sample sizes.

Example 6.5.1. Take $\mathbf{X} = (1,\, X_1,\, X_2)^T$ and the coefficient functions in (6.1.1) are given by

$$a_0(u) = \exp(2u-1), \quad a_1(u) = 8u(1-u), \quad \text{and} \quad a_2(u) = 2\,\sin^2(2\pi u). \qquad (6.5.4)$$

Figure 6.1(a) depicts the marginal distributions for the ratios of the overall RASE defined in (6.5.1), using three bandwidths $h = 0.1$, 0.2 and 0.4. It is evident that the performance of the one-step, the two-step and the fully iterative estimators are comparable for a wide range of bandwidths. As expected, the performance of the two-step estimator is closer to that of the local MLE. Figures 6.1(b)–(d) give the estimate of the coefficient functions from a typical sample. The typical sample is selected in such a way that its RASE-value is the median in the 400 RASE-values. Table 6.1 summarizes the simulation results with μ and σ denoting the mean and standard deviation of the RASE in 400 simulations. Here, ρ_* indicates the correlation coefficient between the RASE of the MLE and the RASE of the one-step (or two-step) method. Note that the correlation coefficients are close to one which indicates that the one-step and two-step methods follow closely the MLE. Note also that the larger the bandwidths, the larger the correlation coefficients. This is due to the fact that a larger bandwidth implies more local data points, which makes the asymptotic theory more relevant. As expected, the correlation coefficients for the two-step method are larger than those of the one-step method, since the former is closer to the MLE.

Table 6.1: Bivariate summary of simulation results for logistic regression model

		MLE		One-step			Two-step		
n	h	μ	σ	μ	σ	ρ_*	μ	σ	ρ_*
400	0.10	2.2278	2.0874	1.8537	0.9759	0.8656	2.1244	1.5315	0.8274
	0.20	1.0669	0.4491	1.0576	0.4378	0.9991	1.0669	0.4491	1.0000
	0.40	0.9454	0.1600	0.9447	0.1593	1.0000	0.9454	0.1600	1.0000
800	0.075	1.2451	0.6639	1.1644	0.3767	0.8342	1.2256	0.5301	0.9656
	0.15	0.7280	0.2573	0.7234	0.2459	0.9993	0.7280	0.2573	1.0000
	0.30	0.7433	0.1009	0.7429	0.1005	1.0000	0.7433	0.1009	1.0000

We now test the accuracy of our standard error formula (6.2.4). The standard deviation, denoted by SD in Table 6.2, of 400 estimated $\hat{a}_j(u_0)$, based on 400 simulations, can be regarded as the true standard errors. The average and the standard deviation of 400 estimated standard errors, denoted by SD_a and SD_{std}, summarize the overall performance of the standard error formula (6.2.4). Table 6.2 presents the results at the points $u_0 = 0.25$, 0.50 and 0.75. It suggests that our standard error formula somewhat underestimates the true standard deviation, though

Table 6.2: Standard deviations of estimators for logistic regression model

			$\hat{a}_0(u)$		$\hat{a}_1(u)$		$\hat{a}_2(u)$	
n	h	u	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})
400	0.2	0.25	0.3185	0.2673 (0.0470)	0.4890	0.4069 (0.0776)	0.5082	0.3986 (0.0893)
		0.50	0.3410	0.2782 (0.0451)	0.5413	0.4330 (0.0809)	0.4135	0.3568 (0.0591)
		0.75	0.4315	0.3542 (0.0776)	0.5372	0.4542 (0.0996)	0.5809	0.4431 (0.0969)
400	0.3	0.25	0.2294	0.2051 (0.0231)	0.3424	0.3201 (0.0447)	0.3317	0.2956 (0.0403)
		0.50	0.2570	0.2315 (0.0315)	0.3931	0.3538 (0.0527)	0.3490	0.3122 (0.0431)
		0.75	0.2850	0.2686 (0.0423)	0.3929	0.3581 (0.0557)	0.3788	0.3328 (0.0500)
800	0.15	0.25	0.2418	0.2214 (0.0214)	0.3638	0.3460 (0.0501)	0.3804	0.3486 (0.0532)
		0.50	0.2249	0.2196 (0.0233)	0.4040	0.3569 (0.0512)	0.3124	0.2812 (0.0356)
		0.75	0.3146	0.2928 (0.0478)	0.4209	0.3804 (0.0667)	0.3987	0.3781 (0.0631)

the difference is within two standard deviations of the Monte Carlo errors. The bias becomes smaller as the number of local data points $n\,h_n$ goes up (see the last two situations). This is consistent with our asymptotic theory.

Next, we conduct a simulation study to see whether the asymptotic null distribution of the test statistic T defined in (6.4.2) depends on the values of $\{a_j\}$ under H_0 (see (6.4.1)) and the limiting conditional null distributions are dependent on the covariate values. To this end, we compute the unconditional null distribution of T with $n = 400$, via 1000 Monte Carlo

simulations, for 5 different sets of values of $\{a_j\}$. These sets of parameters are quite far apart. The resulting 5 densities are depicted in Figure 6.1(e) (thick curves). They are very close, which suggest that the asymptotic null distribution is not very sensitive to the values of $\{a_j\}$. To validate our conditional bootstrap method, five typical data sets were selected from our previous 400 simulations. The estimated conditional bootstrap null distributions, based on 1000 bootstrap samples, are plotted as thin curves in Figure 6.1(e). Six empirical percentiles for five different sets of values of $\{a_j\}$ and covariates are listed in Table 6.3. Both Figure 6.1(e)

Table 6.3: Six empirical percentiles for logistic model

10	25	50	75	90	95
Conditional bootstrap					
7.9579	10.7189	14.2569	18.2625	22.2566	24.9903
8.2450	11.0170	14.6601	18.4897	22.4177	25.5829
8.0004	10.9871	14.2667	18.0413	22.5517	25.1661
8.7738	11.4311	14.8061	18.5209	22.7029	25.3781
8.7906	11.4672	14.9130	18.6168	22.3256	24.7104
Unconditional bootstrap					
7.6381	10.7167	14.5487	18.6276	22.2205	24.4597
7.3478	10.1290	13.9934	17.9622	21.8270	24.4429
7.7238	11.3849	14.6151	18.4796	22.5899	24.7270
8.8042	11.3762	14.8076	18.7571	22.0560	25.1550
8.7865	11.3472	14.5975	18.5198	23.1476	25.8297

and Table 6.3 shows that they are very close to the true null distribution. This demonstrates empirically that our bootstrap method gives a reasonably good approximation to the true null distribution even when the data were generated from an alternative model (6.5.4).

To examine the power of the proposed test, we consider the following null hypothesis

$$H_0 : a_j(u) = \theta_j, \quad j = 0, 1, 2, \quad \text{versus} \quad H_1 : a_j(u) \neq \theta_j, \text{ for at least one } j.$$

The power functions are evaluated under a sequence of the alternative models indexed by β

$$H_1 : a_j(u) = a_{j0} + \beta(a_j^0(u) - a_{j0}), \quad j = 0,\ 1,\ 2 \quad (0 \leq \beta \leq 0.8),$$

where $\{a_j^0(u)\}$ are given in (6.5.4) and $a_{j0} = E\{a_j(U)\}$. Figure 6.1(f) depicts the five power functions based on 1000 simulations for the sample size $n = 400$ at five different significance levels: 0.5, 0.25, 0.10, 0.05, and 0.01. When $\beta = 0$, the special alternative collapses into the null hypothesis. The powers at $\beta = 0$ for the above five significance levels are respectively

0.532, 0.281, 0.101, 0.047 and 0.012. This shows that the conditional bootstrap method gives the right levels of test. The power functions increase rapidly as β increases. This in turn shows that the test proposed in Section 6.4 works well.

6.5.2 Poisson regression

For a Poisson model with the canonical link, by straightforward calculation, the one-step estimator is given similarly to (6.5.2) but now $\mathbf{H}_{n,j} = \sum_{i=1}^n K_h(U_i - u_0)\hat{\lambda}_{i0}(U_i - u_0)^j \mathbf{X}_i \mathbf{X}_i^T$, $j = 0, 1, 2$, $\hat{\lambda}_{i0} = \exp\left[\sum_{j=1}^p \{\hat{a}_{j0} + \hat{b}_{j0}(U_i - u_0)\} X_{ij}\right]$, and $\mathbf{v}_{n,j} = \sum_{i=1}^n K_h(U_i - u_0)(Y_i - \hat{\lambda}_{i0})(U_i - u_0)^j \mathbf{X}_i$, $j = 0, 1$. Using the same arguments as in the previous section, the ridge parameters

$$r_{j,k} = \left(\frac{1}{n}\sum_{i=1}^n X_{ik}^2\right) \hat{\lambda}_0 \, h^{j-1} \int u^j K(u)\, du \qquad \text{with} \qquad \hat{\lambda}_0 = \exp\left(\hat{\mathbf{a}}_0^T \overline{\mathbf{X}}\right) \tag{6.5.5}$$

are employed to alleviate the possible singularity of matrix $\mathbf{H}_{n,j}$ ($j = 0$ and 2) in (6.5.2).

Example 6.5.2. The conditional distribution of Y given covariates U, X_1 and X_2 is taken to be Poisson with the following linear predictor

$$\eta(u, \mathbf{x}) = 5.5 + 0.1\{a_0(u) + a_1(u)\, x_1 + a_2(u)\, x_2\},$$

where the coefficient functions $a_0(u)$, $a_1(u)$ and $a_2(u)$ are the same as those in Example 6.5.1. The coefficients 5.5 and 0.1 are chosen so that the range of simulated data is close to that of the environmental data in Section 6.5.3.

Table 6.4: Bivariate summary of simulation output for Poisson regression model

		MLE		One-step			Two-step		
n	h	μ	σ	μ	σ	ρ_*	μ	σ	ρ_*
200	0.075	0.3632	0.0692	0.3468	0.0562	0.8691	0.3632	0.0692	1.0000
	0.15	0.3220	0.0510	0.3202	0.0504	0.9925	0.3220	0.0510	1.0000
	0.30	0.5852	0.0425	0.5835	0.0426	0.9990	0.5852	0.0425	1.0000
400	0.075	0.2309	0.0352	0.2279	0.0347	0.9866	0.2309	0.0352	1.0000
	0.15	0.2581	0.0325	0.2571	0.0322	0.9942	0.2581	0.0325	1.0000
	0.30	0.5603	0.0292	0.5581	0.0293	0.9988	0.5603	0.0292	1.0000

Figure 6.2 and Table 6.4 summarize the result for $n = 200$. It shows again that the one-step, two-step and the iterative local MLE have comparable performance. A typical estimated

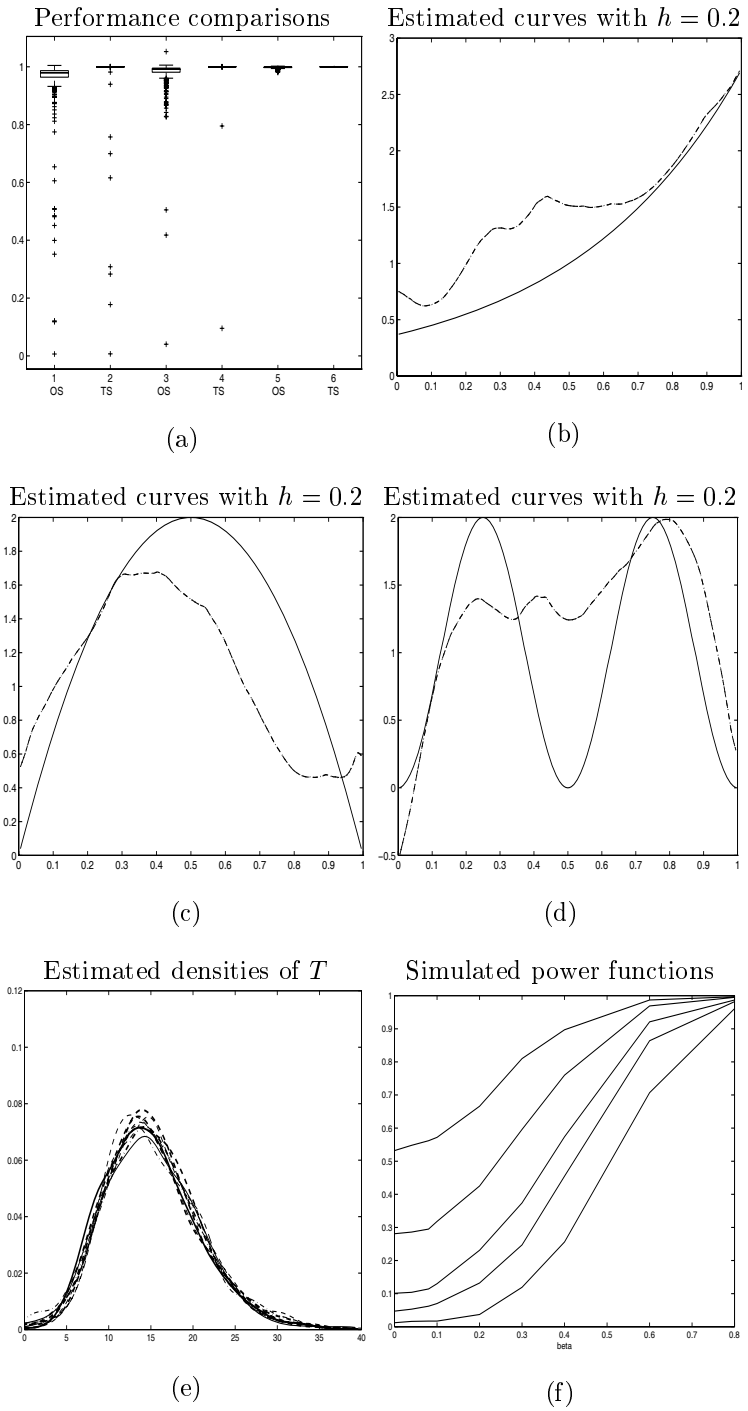


Figure 6.1: *Simulation results for Example 6.5.1 with sample size 400. (a) The boxplots for the ratios of RASE of the one-step and two-step local likelihood approaches to that of the local MLE of* $\mathbf{a}(u)$*, using bandwidths (from left to right)* $h = 0.10$*,* 0.20 *and* 0.40*. (b), (c) and (d) Typical estimates of* $a_0(u)$*,* $a_1(u)$ *and* $a_2(u)$*, respectively, with bandwidth* $h = 0.2$*. Solid curve — true function; dashed curves (from shortest to longest dash) are the one-step, two-step and local MLE, respectively. (e) The estimated densities of* T *for unconditional null distributions (thick curves) and for conditional null distributions (thin curves). (f) The power functions of the test statistic* T*.*

function with bandwidth $h = 0.15$ is presented in Figures 6.2(b)–(d). Because of different noise-to-signal ratios, the functions here are indeed estimated better than those given in Example 6.5.1. Similar to Example 6.5.1, we summarize the performance of our estimated standard error formula (6.2.4) in Table 6.5 Clearly, our estimated standard errors are very close to the

Table 6.5: Standard deviations of estimators for Poisson regression model

			$\hat{a}_0(u)$		$\hat{a}_1(u)$		$\hat{a}_2(u)$	
n	h	u	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})	SD	SD_a (SD_{std})
200	0.15	0.25	0.0105	0.0092 (0.0013)	0.0148	0.0118 (0.0024)	0.0156	0.0126 (0.0026)
		0.50	0.0094	0.0088 (0.0011)	0.0148	0.0112 (0.0022)	0.0150	0.0118 (0.0024)
		0.75	0.0100	0.0088 (0.0011)	0.0142	0.0112 (0.0023)	0.0151	0.0119 (0.0023)
400	0.075	0.25	0.0094	0.0085 (0.0012)	0.0130	0.0106 (0.0021)	0.0136	0.0107 (0.0022)
		0.50	0.0093	0.0083 (0.0011)	0.0127	0.0104 (0.0022)	0.0130	0.0105 (0.0021)
		0.75	0.0090	0.0081 (0.0011)	0.0137	0.0101 (0.0022)	0.0133	0.0102 (0.0022)

true ones.

Similar to Example 6.5.1, the procedure of testing hypothesis is applied to this example. Both unconditional and conditional estimated densities of T are displayed in Figure 6.2(e). Six empirical percentiles are listed in Table 6.6. The corresponding power functions are presented

Table 6.6: Six empirical percentiles for Poisson model

10	25	50	75	90	95
Conditional bootstrap					
12.1646	15.1401	18.6981	22.6260	26.1432	28.8494
11.7506	14.5010	18.0994	22.3809	26.1237	29.4936
11.7946	14.7005	18.3495	22.2918	26.0064	29.2165
11.4662	14.6917	18.2475	22.4623	27.0587	29.6887
11.9894	14.7869	18.5571	22.3593	26.7014	29.7923
Unconditional bootstrap					
11.9492	14.7920	18.5509	22.3383	26.7474	28.8094
11.1599	14.7156	18.7054	22.2915	26.6170	28.9831
11.4378	14.8132	18.4080	22.3890	26.5858	29.4816
11.8238	14.6817	18.5090	22.7050	26.4776	29.3814
11.8365	14.9721	18.7674	22.9402	26.5929	28.9815

in Figure 6.2(f). The same conclusions as those in Example 6.5.1 can be drawn for the Poisson regression model. In particular, the test has the correct levels of significance. See the power functions in Figure 6.2(e) at $\beta = 0$.

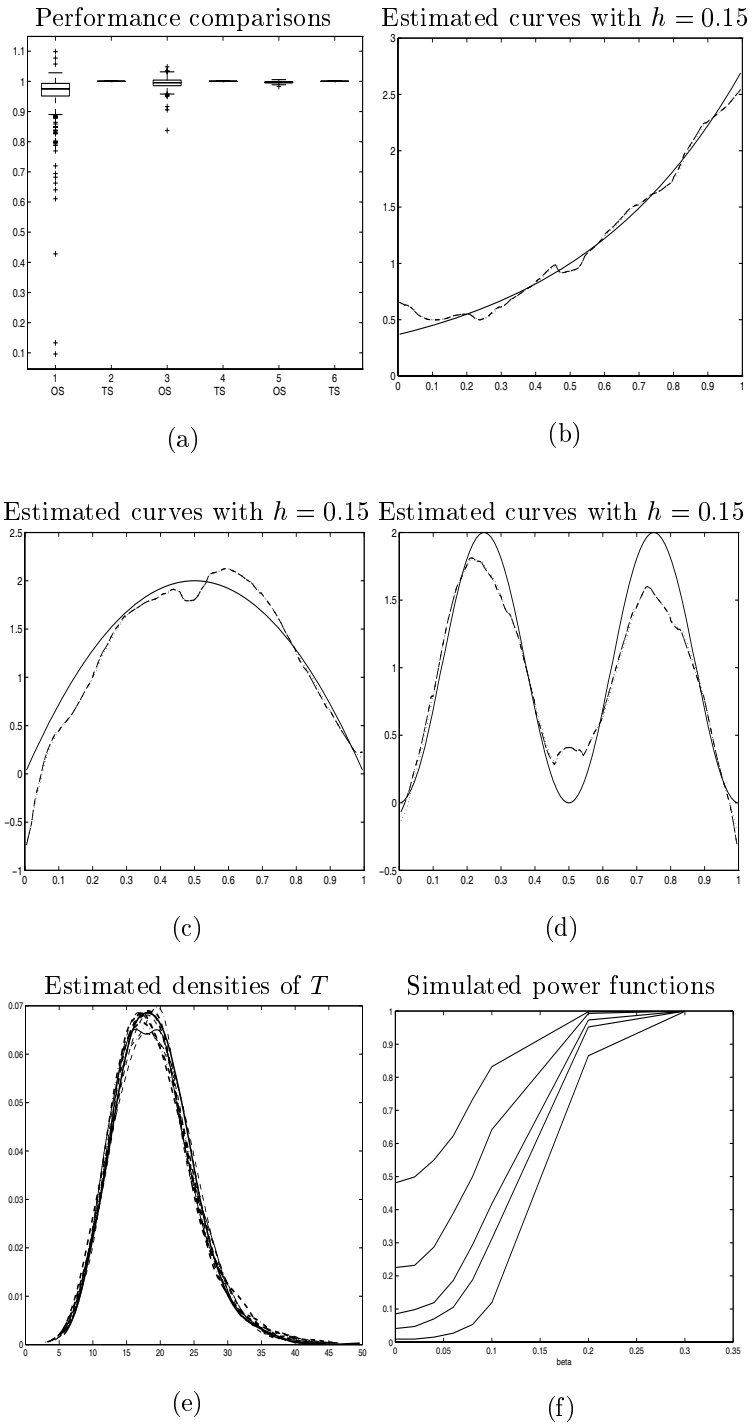


Figure 6.2: *Simulation results for Example 6.5.2 with sample size 200. (a) The boxplots for the ratios of RASE of the one-step and two-step local likelihood approaches to that of the local MLE of* $\mathbf{a}(u)$*, using bandwidths (from left to right)* $h = 0.075$*,* 0.15 *and* 0.30*. (b), (c) and (d) Typical estimates of* $a_0(u)$*,* $a_1(u)$ *and* $a_2(u)$*, respectively, with bandwidth* $h = 0.15$*. Solid curve — true function; dashed curves (from shortest to longest dash) are the one-step, two-step and local MLE, respectively. (e) The estimated densities of* T *for unconditional null distributions (thick curves) and for conditional null distributions (thin curves). (f) The power functions of the test statistic* T*.*

6.5.3 Real-data examples

Example 6.5.3. This example illustrates the proposed procedure via an application to the environmental data set mentioned in the introduction. Of interest is to study the association between levels of pollutants and number of total hospital admissions for circulatory and respiratory problems on every Friday from January 1, 1994 to December 31, 1995 and to examine the extent to which the association varies over time. The covariates are taken as the levels of pollutants sulfur dioxide X_2 (in $\mu g/m^3$), nitrogen dioxide X_3 (in $\mu g/m^3$) and dust X_4 (in $\mu g/m^3$). Since the admissions "events" occur at certain points in time, it is reasonable to model the number of admissions as a Poisson process and use the Poisson regression model with the mean $\lambda(t, \mathbf{x})$ given by

$$\log\{\lambda(t, \mathbf{x})\} = a_1(t) + a_2(t)\, x_2 + a_3(t)\, x_3 + a_4(t)\, x_4. \tag{6.5.6}$$

A multifold cross-validation method is used to select a bandwidth. We partition the data into G groups, — the j^{th} group consisting of data points with indices

$$d_j = \{Gk + j,\ k = 1, 2, \cdots\}, \quad j = 0, \cdots, G-1.$$

For each j, the j-th group of data is deleted and model (6.5.6) is fitted for the remaining data. Then the deviance (McCullagh and Nelder 1989, p.34) or the sum of squares of Pearson's residuals is computed. This leads to two cross-validation criteria

$$CV_1(h) = \sum_{j=0}^{G-1} \sum_{i \in d_j} 2 \left[y_i \log\{y_i / \hat{y}_{-d_j}(U_i, \mathbf{X}_i)\} - \{y_i - \hat{y}_{-d_j}(U_i, \mathbf{X}_i)\} \right],$$

and

$$CV_2(h) = \sum_{j=0}^{G-1} \sum_{i \in d_j} \left\{ \frac{y_i - \hat{y}_{-d_j}(U_i, \mathbf{X}_i)}{\sqrt{\hat{y}_{-d_j}(U_i, \mathbf{X}_i)}} \right\}^2,$$

where $\hat{y}_{-d_j}(U_i, \mathbf{X}_i)$ is a fitted value with the data in d_j deleted. In the implementation, we choose $G = 20$. Figure 6.3(b) depicts the cross-validation functions $CV_1(h)$ and $CV_2(h)$ which give the optimal bandwidth $h = 0.1440 \times 105$. To see how sensitive the above partition is to the CV curves, the data set is randomly partitioned into 20 groups, and then cross-validation scores are computed based on the same procedure described above. The results are depicted in Figure 6.3(c), which, in conjunction with Figure 6.3(b), shows that the cross-validation functions is not very sensitive to the partition. The bandwidth $h = 0.1440 \times 105$ results in undersmoothed

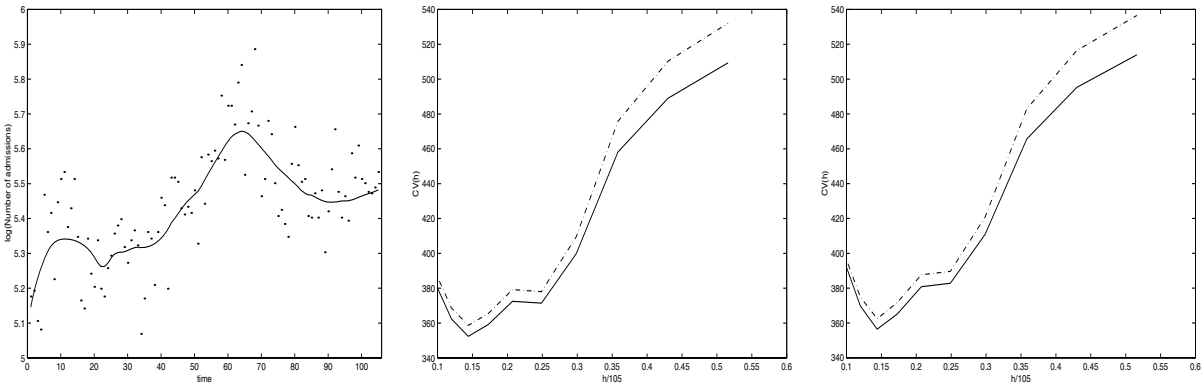
estimated coefficients. From Figure 6.3(c), it can be seen that there is a local minimum around $h = 0.25 * 105$ of the cross-validation score curve. Using bandwidth $h = 0.2488 \times 105$, the estimated coefficient functions based on the one-step procedure are summarized in Figure 6.4 since the results based on both the one-step and the fully iterative methods are very close. They describe the extent to which the association between the pollutants and the number of hospital admissions vary over time. The figure shows clearly that the coefficient functions vary with time. The two dashed curves are the estimated function plus/minus twice the estimated standard errors. They give us an idea of the pointwise confidence intervals with bias ignored.

A question arises whether or not the data are highly correlated. To check for the serial correlation, Pearson's residuals are computed. The time series plot of the residuals is given in Figure 6.5(a) and the plot of the corresponding autocorrelation coefficients against time lag is presented in Figure 6.5(b). There is no pattern in Figure 6.5(a). Thus, Figure 6.5(a) together with Figure 6.5(b) lead to the conclusion that there is no evidence that the data are serially correlated.

We now apply the procedure proposed in Section 6.4 to testing whether the coefficients are actually time varying. The MLE under the null hypothesis is $(5.4499,\ -0.0025,\ 0.0015,\ -0.0005)$ with an estimated standard deviation $(0.0195,\ 0.0006,\ 0.0006,\ 0.0005)$. The test statistic (6.4.2) is $T = 389.41$. Based on 1000 bootstrap replications, the sample mean and sample variance of T^* are 26.64 and 48.40, respectively. The distribution of T is approximated by a χ^2 distribution with degrees of freedom 27 (see Figure 6.6). The p-value is close to zero, which strongly rejects the null hypothesis. Therefore, it suggests that the varying-coefficient model gives a much better fit than the parametric model.

Now we use our testing approach proposed in Section 6.4 to check whether there is any covariate that can be deleted from the model. We start with X_4 since the parametric Poisson model concludes that the dust level (X_4) is not statistically significant. To examine if the variable X_4 is significant in the varying-coefficient model, we apply the idea in Section 6.4 to testing the hypothesis: the function $a_4(\cdot)$ is zero. The maximum likelihood ratio test statistic is $T = 20.1847$. Based on 1000 bootstrap samples, the p-value is 0.321 (the sample mean and variance of T^* are 17.7352 and 37.1976, respectively). Therefore, the variable X_4 can be dropped from the varying-coefficient model. After deleting the variable dust level (X_4), we apply the same procedure as above to test whether X_3 is statistically significant in the varying-coefficient model. That is to test $H_0 : \log\{\lambda(t, \mathbf{x})\} = a_1(t) + a_2(t)\, x_2$ against

$H_1 : \log\{\lambda(t,\, \mathbf{x})\} = a_1(t) + a_2(t)\, x_2 + a_3(t)\, x_3$. As a result, the maximum likelihood ratio test statistic is $T = 39.7473$ and the p-value is 0.039 (the sample mean and variance of T^* are 27.5071 and 39.5808, respectively), based on 1000 bootstrap samples. Therefore, the variable nitrogen dioxide (X_3) is significant at the significant level 0.05. By the same token, the variable sulfur dioxide (X_2) is significant too.



(a) (b) (c)

Figure 6.3: *(a) The scatterplot of log transformation of environmental data set studied in Example 6.5.3. The curve is the estimate of* $a_1(t) + a_2(t)\,\bar{x}_1 + a_3(t)\,\bar{x}_2 + a_4(t)\,\bar{x}_3$, *where* $\bar{x}_j$ *is the average pollutant level* x_j. *(b) The plot of the cross-validation functions* $CV_1(h)$ *(solid line) and* $CV_2(h)$ *(dashdot line) against bandwidth. (c) The same as those in (b), but the cross-validation is based on random partitions of the data set.*

Example 6.5.4. Now we apply the methodology proposed in this paper to analyze the data set: *Burns data*, collected by General Hospital Burn Center at the University of Southern California. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise, and covariates X_1=*age*, X_2=*sex*, $X_3 = \log(\text{burn area}+1)$ and binary variable X_4 =*Oxygen* (0 if oxygen supply is normal, 1 otherwise) are considered. We are interested in studying how burn areas and the other variables affect survival probabilities for victims at different age groups. This naturally leads to the following varying-coefficient model

$$\text{logit}\{p(x_1, x_2, x_3, x_4)\} = a_1(x_1) + a_2(x_1)\, x_2 + a_3(x_1)\, x_3 + a_4(x_1)\, x_4. \tag{6.5.7}$$

Figure 6.7 presents the estimated coefficients for model (6.5.7) via the one-step approach with bandwidth $h = 65.7882$, selected by a cross-validation method.

A natural question arises whether the coefficients in (6.5.7) are actually varying. To see this, we consider the parametric logistic regression model

$$\text{logit}\{p(x_1,\, x_2,\, x_3,\, x_4)\} = \beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \beta_4\, x_4 \tag{6.5.8}$$

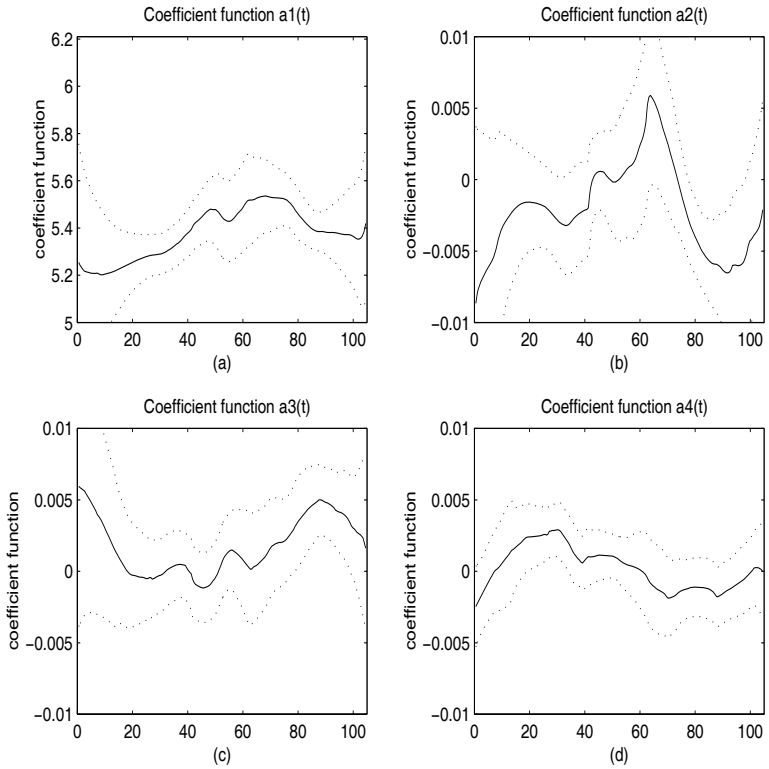
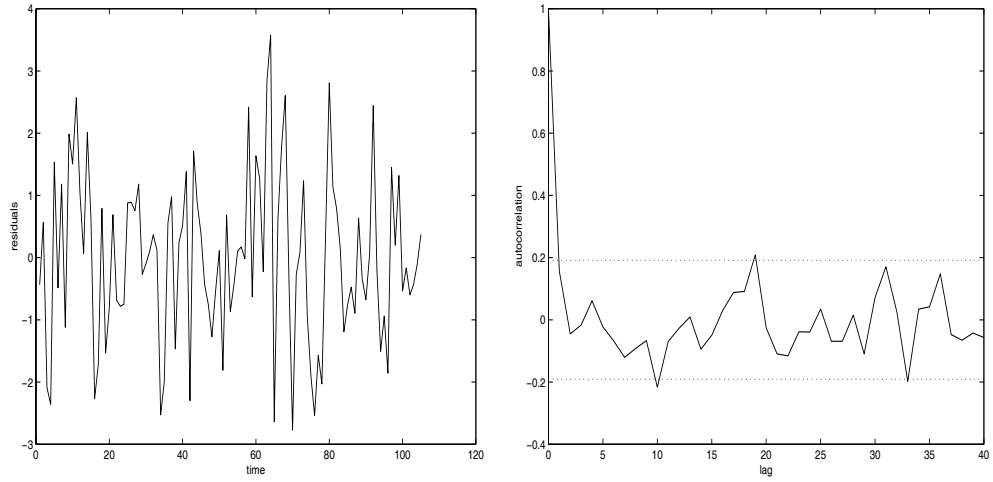


Figure 6.4: *The estimated coefficient functions via the one-step approach with bandwidth chosen by the CV. The dashed curves are the estimated function plus/minus twice estimated standard errors.*



(a) (b)

Figure 6.5: *(a) The time series plot of Pearson's residuals. (b) The plot of the autocorrelation coefficients versus time lag. The two dashed curves are* $\pm 1.96/\sqrt{n}$*, where* n *is the sample size.*

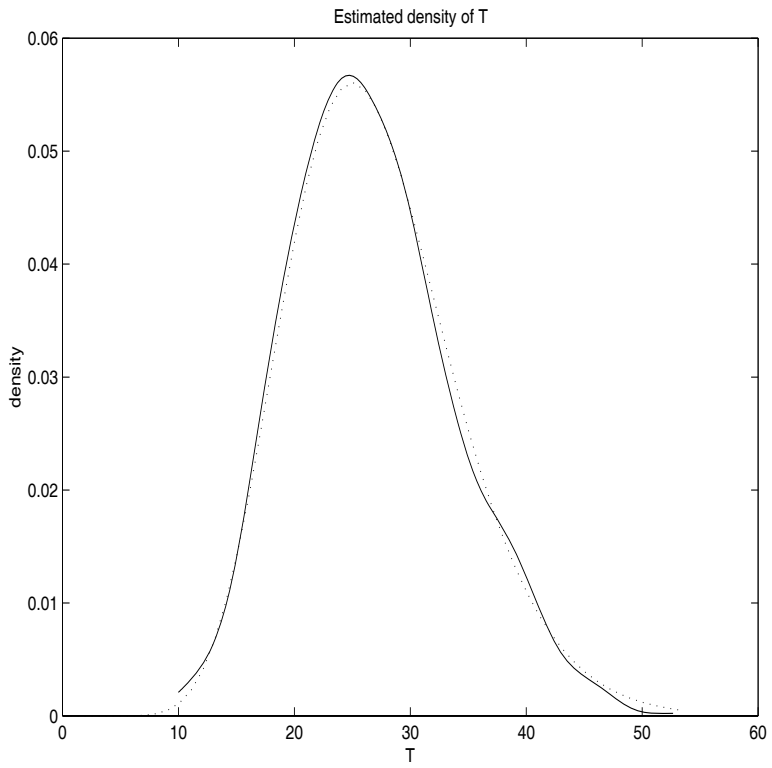


Figure 6.6: *The estimated density of T by Monte Carlo simulation. The solid curve is the estimated density, and the dashed curve stands for the density of chi-squared distribution with degrees of freedom 27.*

as the null model. As a result, the MLE of $(\beta_0, \cdots, \beta_4)$ in model (6.5.8) and its standard deviation are (23.221, -6.149, -0.466, -2.450, -0.968) and (1.918, 0.665, 0.283, 0.221, 0.290), respectively. The test statistic T proposed in Section 6.4 is 54.9601 with p-value 0.000, based on 1000 bootstrap samples (the sample mean and variance of T^* are 5.9756 and 10.7098, respectively). This implies that the varying-coefficient logistic regression model fits the data much better than the parametric fit. It also allows us to examine the extent to which the regression coefficients vary over different ages.

To examine whether there is any gender gap for different age groups or if the variable X_4 affects the survival probabilities for different age of burn victims, we consider testing hypothesis H_0 : both $a_2(\cdot)$ and $a_4(\cdot)$ are constant under model (6.5.7). The corresponding test statistic T is 3.2683 with p-value 0.7050, based on 1000 bootstrap samples. This in turn suggests that the coefficient functions $a_2(\cdot)$ and $a_4(\cdot)$ are independent of age and indicates that there are no gender differences for different age groups.

Finally, we examine whether both covariates *sex* and *Oxygen* are statistically significant in model (6.5.7). The likelihood ratio test for this problem is $T = 11.2727$ with p-value 0.0860, based on 1000 bootstrap samples (the sample mean and variance of T^* are 5.2867 and 9.7630, respectively). Both covariates *sex* and *Oxygen* are not significant at level 0.05. This suggests that gender and oxygen do not play a significant role in determining the survival probability of a victim.

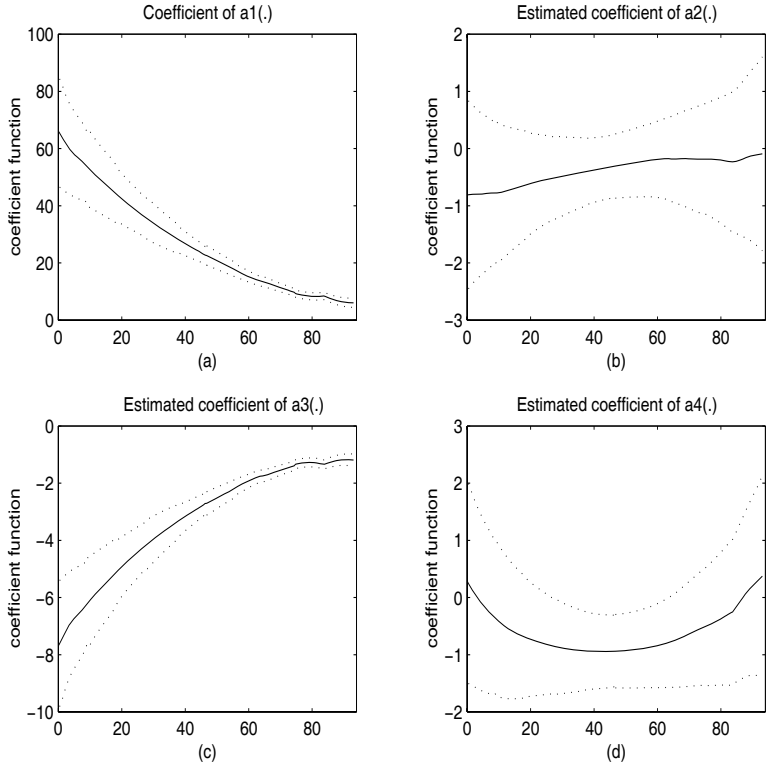


Figure 6.7: *The estimated coefficient functions (the solid curves) via one-step approach with bandwidth chosen by the CV. The dot curves are the estimated functions plus/minus twice estimated standard errors.*

Chapter 7

SiZer Map Based on Local Likelihood

7.1 Introduction

Consider the bivariate data $(X_1, Y_1), \cdots, (X_n, Y_n)$, an i.i.d. sample from the model

$$Y = m(X) + \varepsilon,$$

where ε is random error with $E(\varepsilon|X) = 0$, $\varepsilon^2 = \sigma^2(X)$. As mentioned in Chapter 5, local polynomial regression can be employed to estimate the regression function. The choice of bandwidth h is of crucial importance in nonparametric kernel smoothing regression. If h is chosen too large then the resulting estimate misses fine features of the regression curve, while if h is selected too small then spurious sharp structure become visible. There are a large number of proposals on bandwidth selection in literature. An excellent survey of bandwidth selection procedures in kernel density estimation can be found in Jones, Marron and Sheather (1996a, b). Chapter 4 of Fan and Gijbels (1996) outlined some approaches of bandwidth selection in the context of local polynomial regression. A "best" data-driven bandwidth for local polynomial regression can be constructed using the ideas of cross-validation (see, for example, Härdle, 1990), nearest neighbor bandwidth (Fan and Gijbels, 1995), and plug-in (Gasser, Kneip, Köhler, 1991, Sheather and Jones, 1991 and Ruppert, Sheather and Wand, 1995). All of these methods try to find a single best bandwidth under some criteria. Chaudhuri and Marron (1999) proposed the idea of family smoothing and SiZer map from scale-space point of view. Scale-space ideas from computer vision provide a different view-point on kernel smoothing.

The "scale space surface," the family of all kernel smooths indexed by the bandwidth h, is a model used in computer vision. The essential idea is that large h models macroscopic (distant) vision where only large-scale features can be resolved, and small h models microscopic (zoomed in) resolution of small-scale features. This is very different from the classical statistical approach, where the focus is the underlying regression function m. See Chaudhuri and Marron (1999) for detailed discussion.

Chaudhuri and Marron (1999) have developed the SiZer map in least squares settings from the point of view of scale space. The SiZer map is powerful for finding which features in noisy data are strong enough to be distinguished from background noise. However this approach may be inefficient for discrete data when the likelihood function or the quasi-likelihood function of data is available. This is seen in Figures 7.1 and 7.2. The color scheme and line type of the SiZer map developed in this chapter will exactly follow those of Chaudhuri and Marron (1999). Specifically, behavior at an (x, h) location is presented via the SiZer color map where blue (black in versions where only black and white are available) indicates locations where the mean function is significantly increasing, red (white in black-white versions) shows where it is significantly decreasing, and purple (gray in black-white versions) indicates where the mean function is not significantly different from zero. Moreover, a location is shaded gray when the "effective sample size in the window" is less than 5. The dotted curves in the SiZer maps show "effective window widths" of the smoothing windows, as intervals representing $\pm 2h$ (i.e. $\pm 2h$ standard deviations of the Gaussian kernel). A reference bandwidth is highlighted by the horizontal bar.

Example 7.1.1 (*Poisson regression*) In this example, the covariate X is from an equally spaced design on $[0,\ 10\pi]$, and the conditional distribution of Y given X is a Poisson distribution with mean function

$$\lambda(x) = \exp\left\{\frac{15\sin(x)}{x+4}\right\}.$$

Figure 7.1 depicts the plot of the mean function. Figures 7.2 and 7.3 compare the SiZer map proposed by Chaudhuri and Marron (1999) and the SiZer map based on local likelihood for sample sizes 500 and 200, respectively. In Figure 7.2, the two kinds of SiZer maps look similar because the sample size is large so all of significant features are distinguished from background noise. However, Figure 7.3 indicates that the SiZer map based on local likelihood is more efficient in distinguishing important features than the original SiZer

This chapter will develop an extension of the SiZer map based on local likelihood. As pointed out by Chaudhuri and Marron (1999), the extension of the SiZer map to the context

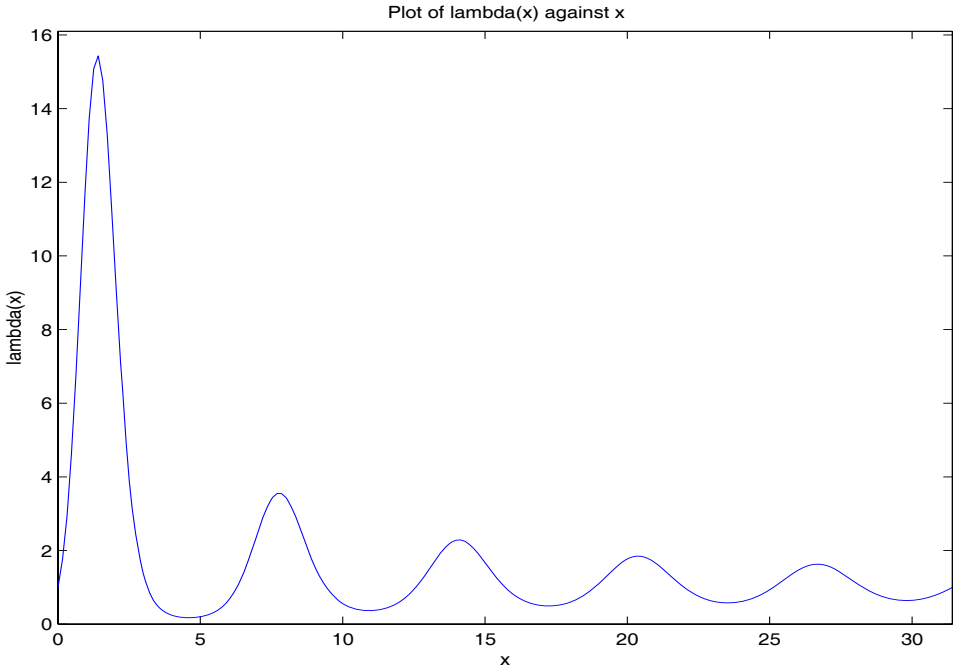


Figure 7.1: *True signal in Example 7.1.1.*

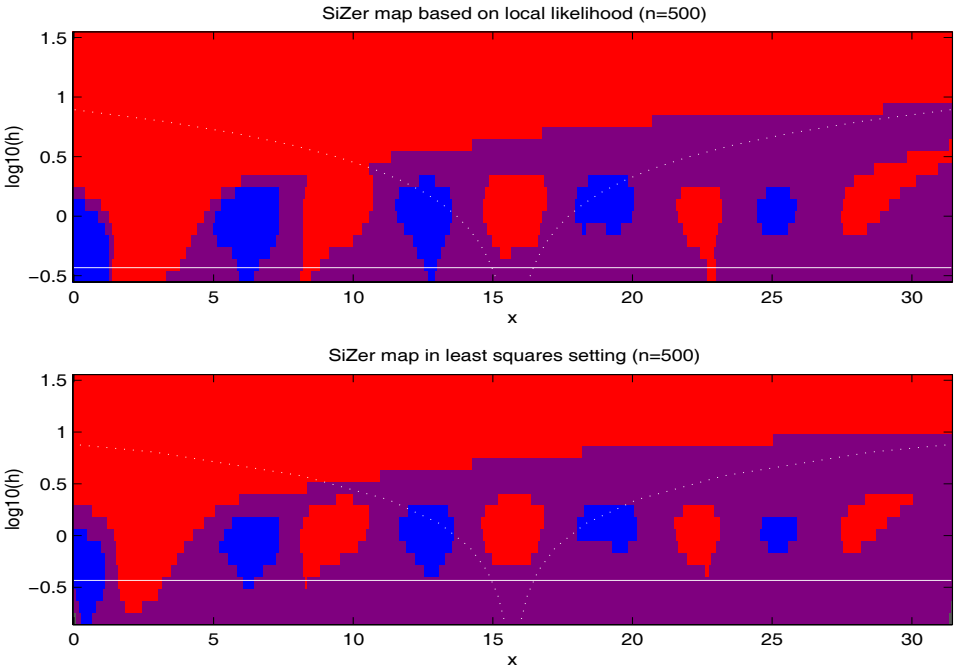


Figure 7.2: *SiZer maps for sample size* $n = 500$. *The top panel is the SiZer map based on local likelihood, and the bottom panel is the SiZer map proposed by Chaudhuri and Marron (1999).*

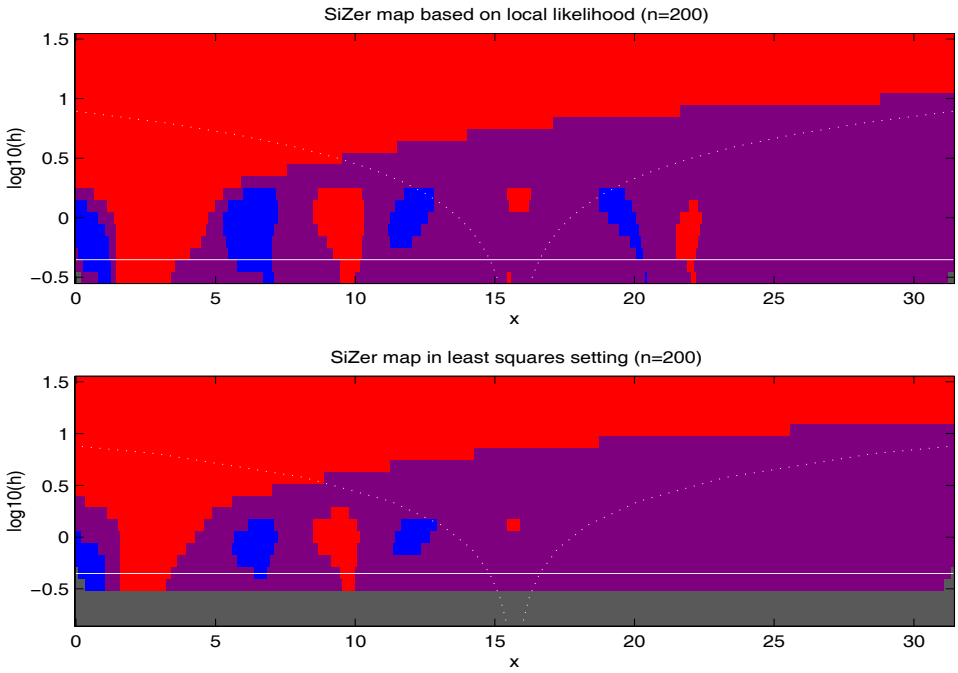


Figure 7.3: *SiZer maps for sample size* $n = 200$. *The top panel is the SiZer map based on local likelihood, and the bottom panel is the SiZer map proposed by Chaudhuri and Marron (1999).*

of local likelihood may be straightforward. However, unlike the least-squares setting, the solution for local likelihood score equations generally does not have a closed form. To obtain the solution, one usually uses an iterative algorithm, such as Newton-Raphson. The computational cost of the iterative method can be very expensive as one needs to maximize the local quasi-likelihood function defined below for many distinct values of x in order to estimate the whole curve over all of the scale. Some computational implementation issues will be disscussed in this chapter. Some applications for real data examples are also given.

7.2 Generalized Linear Models and Quasi-likelihood Functions

In this chapter, definitions of generalized linear models and quasi-likelihood functions follow those in McCullagh and Nelder (1989). Suppose that $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ are iid samples from the population $(\mathbf{X}, Y)$, where $\mathbf{X}$ is a d-dimensional real vector of covariates, and Y is a scalar response variable. The conditional density of Y given covariate $\mathbf{X} = \mathbf{x}$ belongs to the canonical exponential family:

$$f_{Y|\mathbf{X}}(y|\mathbf{x}) = \exp\{[\theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}]/a(\phi) + c(y, \phi)\} \tag{7.2.1}$$

for known functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$. In parametric generalized linear models it is usual to model a transformation of a regression function $m(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x})$ as linear, that is

$$\eta(\mathbf{x}) = g\{m(\mathbf{x})\} = \mathbf{x}^T\boldsymbol{\beta},$$

and g is a known *link* function. If $g = (b')^{-1}$, then g is called the canonical link because $b'\{\theta(\mathbf{x})\} = m(\mathbf{x})$.

Under model (7.2.1), it can be easily shown that the conditional mean and conditional variance are given respectively by $m(\mathbf{x}) = E(Y|\mathbf{X}=\mathbf{x}) = b'\{\theta(\mathbf{x})\}$, and $\mathrm{var}(Y|\mathbf{X}=\mathbf{x}) = a(\phi)b''\{\theta(\mathbf{x})\}$. Since our primary interest is to estimate the mean function, without loss of generality, the factors related to the dispersion parameter ϕ are omitted. This leads to the following conditional log-likelihood function

$$\ell\{\theta, y\} = \theta(\mathbf{x})y - b\{\theta(\mathbf{x})\}.$$

There are many practical circumstances in which the conditional likelihood of Y is unknown, one only knows the relationship between the mean function and the variance function. In this situation estimation of the mean function can be achieved by replacing the conditional log-likelihood $\log\{f_{Y|\mathbf{X}}(y|\mathbf{x})\}$ by a quasi-likelihood function $Q\{m(\mathbf{x}),)\}$. If the conditional variance is modeled as $\mathrm{var}(Y|\mathbf{X}=\mathbf{x}) = V\{m(\mathbf{x})\}$ for some known positive function V, then the corresponding quasi-likelihood $Q(\mu, y)$ satisfies

$$\frac{\partial}{\partial\mu}Q(\mu, y) = \frac{y-\mu}{V(\mu)} \tag{7.2.2}$$

(due to Wedderburn, 1974). The quasi-score (7.2.2) possesses properties similar to those of the usual log-likelihood score function. Quasi-likelihood methods behave analogously to the usual likelihood methods and thus are reasonable substitutes when the likelihood function is not available. Note that the log-likelihood of the one-parameter exponential family is a special case of a quasi-likelihood function with $V = b'' \circ (b')^{-1}$, where $\circ$ denotes function composition.

7.3 Local Quasi-likelihood

Since SiZer map for multi-dimensional covariates is out of the scope of this chapter, we only consider the $\mathbf{x}$ in (7.2.1) as one-dimensional. Fan, Heckman and Wand (1995) showed that the local polynomial quasi-likelihood method inherits the good statistical properties of the local polynomial least-squares approach to smoothing. Because of the quasi-likelihood generality,

we will present our result based on quasi-likelihood in this section. Results for the exponential family and for generalized linear models follow as a special case.

Suppose that the second derivative of the $\eta(x)$ exists and is continuous. For each given point x_0, we approximate the function $\eta(x)$ locally by a linear function $\eta(x) \approx \beta_0 + \beta_1(x - x_0)$ for x in a neighborhood of x_0. Note that β_0 and β_1 depend on x_0. Based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$, the local quasi-likelihood is

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q[g^{-1}\{\beta_0 + \beta_1(X_i - x_0)\}, Y_i]K_h(X_i - x_0), \tag{7.3.3}$$

where $K_h(\cdot) = K(./h)/h$ with $K(\cdot)$ being a kernel function, $h = h_n > 0$ is a bandwidth. Define the local quasi-likelihood estimator of $\boldsymbol{\beta}$ to be

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta} \in R^2} Q(\boldsymbol{\beta}). \tag{7.3.4}$$

Thus the local linear quasi-likelihood estimator of $\eta(x)$ is given by

$$\hat{\eta}(x) = \hat{\beta}_0,$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$. The conditional mean function $m(x)$ can then be estimated by applying the inverse of the link function to give

$$\hat{m}(x) = g^{-1}\{\hat{\eta}(x)\}.$$

7.4 SiZer Map Based on Local Quasi-likelihood

The SiZer map developed here is similar to that of Chaudhuri and Marron (1999). It is based on estimation of first order derivatives of the estimated function. Because the link function $g(\cdot)$ usually is a strictly monotone function, we may study the significance features of $\eta(x)$ instead of the mean function $m(x) = g^{-1}\{\eta(x)\}$. Our approach to the visual assessment of significance of features such as peaks and valleys in a family of smoothers $\{\hat{\eta}_h(x) : h \in [h_{\min}, h_{\max}]\}$ is based on confidence limits for the derivative in scale space $\eta'_h(x)$. The range of bandwidths will be discussed later on.

7.4.1 One-step local quasi-likelihood estimator

Repeated calculation of smoothers is required for such color maps. Unlike the least-squares setting, the solution for (7.3.4) generally does not have a closed form. To obtain the solution,

one usually uses an iterative algorithm, such as Newton-Raphson. The computational cost of the iterative method can be very expensive as one needs to maximize the local quasi-likelihood (7.3.3) for many distinct values of x in order to obtain the function $\hat{\eta}'(\cdot)$. To reduce the computational cost, we suggest replacing the iterative local quasi-likelihood estimator by an explicit non-iterative estimator. An excellent candidate is the one-step Newton-Raphson scheme, which has been frequently used in parametric models (see for example, Bickel 1975) and extended to the setting of quasi-likelihood recently by Fan and Chen (1999). Since the local quasi-likelihood method involves finding hundreds of parametric maximum likelihood estimates, the computational gain of one-step local quasi-likelihood estimates is much more significant than that for parametric models. It can be shown that the one-step local quasi-likelihood estimate does not lose any statistical efficiency provided that the initial estimator is good enough (See Fan and Chen (1999) for the univariate case, and Chapter 6 for the multivariate case).

Now let us describe the one-step local quasi-likelihood estimator. Let $Q'(\boldsymbol{\beta})$ and $Q''(\boldsymbol{\beta})$ be the gradient and Hessian matrix of the local quasi-likelihood $Q(\boldsymbol{\beta})$. Given an initial estimator $\hat{\boldsymbol{\beta}}_0(x_0) = (\hat{\beta}_0(x_0), \hat{\beta}_1(x_0))^T$, the Newton-Raphson algorithm finds an updated estimator

$$\hat{\boldsymbol{\beta}}_{\mathrm{OS}} = \hat{\boldsymbol{\beta}}_0(x_0) - [Q''\{\hat{\boldsymbol{\beta}}_0(x_0)\}]^{-1} Q'\{\hat{\boldsymbol{\beta}}_0(x_0)\}. \tag{7.4.5}$$

This one-step estimator inherits clearly the computation expediency of least-squares local polynomial fitting.

An important issue here is how to choose an good initial value. Fan and Chen (1999) suggested the use of a local least-squares estimator with some modification as initial value. One may also follow the proposal in Section 5.2.4. When one constructs the SiZer map, it is necessary to calculate local quasi-likelihood estimates with different bandwidths. Therefore the local quasi-likelihood estimates based on a larger bandwidth may serve as a natural candidate for an initial estimate of the local one-step quasi-likelihood estimator based on the next smaller bandwidth. Take $h_{\max}$ to be the range of the data, then the starting estimator will be based on a parametric model. Figures 7.4 and 7.5 show that the latter approach to starting values is more efficient in distinguishing important features.

7.4.2 Numerical implementation of binned methods

The one-step local quasi-likelihood estimator can save computational cost by a factor of tens without deteriotating the performance of the fully iterative local quasi-likelihood estimator. In least-squares setting, binned methods can save computational time by a factor of hundreds. The idea discussed by Fan and Marron (1994) can be directly extended to the setting of local quasi-likelihood. As aforementioned, this chapter only considers that the link function g is a canonical link. Extension to other link functions does not involve any difficulty except some additional tedious notation. Therefore it follows that

$$Q'(\boldsymbol{\beta}) = \sum_{i=1}^{n} [Y_i - g^{-1}\{\beta_0 + \beta_1(X_i - x_0)\}] K_h(X_i - x_0) \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix},$$

and

$$Q''(\boldsymbol{\beta}) = -\sum_{i=1}^{n} V\{\beta_0 + \beta_1(X_i - x_0)\} K_h(X_i - x_0) \begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix} (1, X_i - x_0)$$

where $V(u) = dg^{-1}(u)/du$, the variance function of $V(Y|X)$, which is always positive. Therefore the Hessian matrix $Q''(\boldsymbol{\beta})$ is always positive definite and $Q(\boldsymbol{\beta})$ is a convex function with respect to $\boldsymbol{\beta}$.

In this chapter, we always use the "linear binning" described in Fan and Marron (1994). For the equally spaced grid of points $\{x_j : j = 1, \cdots, g\}$, the sample $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ is summarized by the binned data

$$\{(x_j, \bar{Y}_j, c_j) : \ j = 1, \cdots, g\}.$$

Using the "bin averages" $\{\bar{Y}_j\}$ and the "bin counts" $\{c_j\}$ given by

$$\bar{Y}_j \equiv \text{avg}\{Y_i : \ i \in I_j\}, \quad \text{and} \quad c_j = \#(I_j)$$

where the I_j are the index sets

$$I_j \equiv \{i : X_i \to x_j\}, j = 1, \cdots, g$$

Denote $\kappa_{l,j} = K_h(j\Delta)(j\Delta)^l$, where Δ is the binwidth and for $l = 0,\ 1,\ 2$ set

$$\begin{aligned} U_l(x_{j'}) &= \sum_{j=1}^{g} \kappa_{l,j-j'} c_j \bar{Y}_j - \sum_{j=1}^{g} \kappa_{l,j-j'} g^{-1}\{\beta_0 + \beta_1(j - j')\Delta\} c_j, \\ h_l(x_{j'}) &= \sum_{j=1}^{g} \kappa_{l,j-j'} c_j V\{\beta_0 + \beta_1(j - j')\Delta\} \end{aligned}$$

and

$$\begin{aligned}\mathbf{U}(x_{j'}) &= (U_1(x_{j'}), U_2(x_{j'})^T \\ \mathbf{H}(x_{j'}) &= \begin{pmatrix} h_0(x_{j'}) & h_1(x_{j'}) \\ h_1(x_{j'}) & h_2(x_{j'}) \end{pmatrix}.\end{aligned}$$

Then

$$\boldsymbol{\beta}_{\mathrm{OS}}(x_{j'}) = \hat{\boldsymbol{\beta}}_0(x_{j'}) - [\mathbf{H}(x_{j'})]^{-1}\mathbf{U}(x_{j'}).$$

A natural estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}$ is the corresponding sandwich formula. That is

$$\mathrm{Cov}\{\hat{\boldsymbol{\beta}}(x_{j'})\} = \mathbf{H}^{-1}(x_{j'})\mathbf{V}(x_{j'})\mathbf{H}^{-1}(x_{j'}),$$

where

$$\mathbf{V}(x_0) = \sum_{i=1}^{n} V\{\beta_0 + \beta_1(X_i - x_0)\}K_h^2(X_i - x_0)\begin{pmatrix} 1 \\ X_i - x_0 \end{pmatrix}(1, X_i - x_0).$$

Define $\tau_{l,j} = K_h^2(j\Delta)(j\Delta)^l V\{\beta_0 + \beta_1(j - j')\Delta\}$ for $l = 0,\ 1,\ 2$. Then the covariance matrix of the one-step estimator at grid point $x_{j'}$ is

$$\mathbf{V}(x_{j'}) = \begin{pmatrix} \tau_{0,j} & \tau_{1,j} \\ \tau_{1,j} & \tau_{2,j} \end{pmatrix}$$

Thus we can calculate the covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}$.

7.4.3 SiZer map

It is essential to the SiZer map to construct a good simultaneous confidence interval for the estimated derivative. Confidence limits for $\eta_h'(x)$ are of the form

$$\hat{\eta}_h'(x) \pm q \cdot \hat{sd}\{\hat{\eta}_h'(x)\}, \tag{7.4.6}$$

where q is an appropriate quantile, and the standard deviation is estimated as discussed in Section 7.4.2. The construction of simultaneous confidence interval relates to the topic of multiple comparisons in classical statistics. Chaudhuri and Marron (1999) proposed and compared several candidates for the quantile q. Intuitively, the pointwise Gaussian quantile $\Phi^{-1}(1 - \alpha/2)$ is too small, meaning the length of the confidence interval is too short, which leads to too many features being flagged as "significant". The calculation of the quantile q based on bootstrap methods is time-consuming. Here is a brief description how to construct time-saving

and appropriate confidence limits for $\eta_h'(x)$, whose behaviour is similar to that constructed via bootstrap. See Chaudhuri and Marron (1999).

An (x, h) location (in scale space) is called significantly increasing, decreasing, or not significant, where zero is below, above or within these confidence limits respectively. In this paper we take q as an approximate by simultaneous over x Gaussian quantiles, based on "number of independent blocks". The quantile q is based on the fact that when x and x' are sufficiently far apart, so that the kernel windows centered at x and x' are essentially independent, but when x and x' are close together, the estimates are highly correlated. The simultaneous confidence limit problem is then approximated by m independent confidence interval problems, where m reflects the " number of independent blocks". We estimated m through an "estimated effective sample size", defined for each (x, h) as

$$\mathrm{ESS}(x, h) = \frac{\sum_{i=1}^{n} K_h(X_i - x)}{K_h(0)}.$$

Note that when $K(\cdot)$ is a uniform kernel, $\mathrm{ESS}(x, h)$ is the number of data points in the kernel window centered at x. For other kernel shapes, points are downweighted according to the height of the kernel function, just as they are in the average represented by the kernel estimators. Next we choose m to be essentially the number of "independent blocks of average size available from our data set of size n"

$$m(h) = \frac{n}{\mathrm{avg}_x \mathrm{ESS}(x, h)}.$$

Now assuming independence of these $m(h)$ blocks of data the approximate simultaneous quantile is

$$q_\alpha(h) = \Phi^{-1}\left\{\frac{1 + (1-\alpha)^{1/m(h)}}{2}\right\}. \tag{7.4.7}$$

The above quantile will be used in this chapter. The quantity ESS is also useful to highlight regions where the normal approximation implicit in (7.4.6) could be inadequate. This plays a role similar to np in the Gaussian approximation to the binomial. So regions where $\mathrm{ESS}(x, h) < n_0$ (we have followed the standard practice of $n_0 = 5$ at all points here) are shaded gray in our SiZer map, to rule out spurious features, and also to indicate regions where the smooth is essentially based on data. The above calculation of the block size $m(h)$ is modified to avoid problems with small ESS as

$$m(h) = \frac{n}{\mathrm{avg}_{x \in D_h} \mathrm{ESS}(x, h)},$$

where D_h is the set of x locations where the data are dense

$$D_h = \{x : \mathrm{ESS}(x, h) \geq n_0\}.$$

Bandwidth selection is not an important issue for the SiZer map because it is based on the idea of family smoothing. However, a reference bandwidth may help one to interpret the map. Fan and Chen (1999) have made a simple connection for the bandwidth selection problem between the local least-squares method and the local quasi-likelihood method. They suggested use of an estimated optimal bandwidth for the least-squares local polynomial estimator with some modification as the bandwidth for the local one-step quasi-likelihood estimator. Thus the Ruppert-Sheather-Wand direct plug in bandwidth is taken as a reference bandwidth, highlighted as a horizontal bar in SiZer map and as a thick red curve in family plot.

The bandwidth range $[h_{\min}, h_{\max}]$ can be chosen in several ways. In this chapter $h_{\min}$ is taken to be the smallest bandwidth for which there is no substantial distortion in construction of the binned implementation of the smoother, $h_{\min} = 5 * (\text{binwidth})$, and $h_{\max}$ to be the range of the data.

7.5 Simulation and Application

In this section, the proposed SiZer map is illustrated with both simulation examples and a real data example. It will be shown that the SiZer map based on the proposed one-step estimator behaves as well as the one based on the maximum local quasi-likelihood estimate with full iterations.

7.5.1 Poisson regression

For a Poisson regression model, the conditional distribution of Y given X is Poisson mean function $\lambda(x)$. The canonical link for Poisson regression is log-link. With the canonical link, the local (conditional) likelihood, based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$ is

$$Q\{\boldsymbol{\beta}(x_0)\} = \sum_{i=1}^{n} \{Y_i \mathbf{X}_i^T \boldsymbol{\beta}(x_0) - \exp(\mathbf{X}_i^T \boldsymbol{\beta}(x_0)\} K_h(X_i - x_0),$$

where $\mathbf{X}_i = (1, X_i - x_0)^T$ and $\boldsymbol{\beta}(x_0) = (\beta_0(x_0), \beta_1(x_0))^T$. Therefore

$$Q'(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i \{Y_i - \exp(\mathbf{X}_i^T \boldsymbol{\beta})\} K_h(X_i - x_0),$$

and

$$Q''(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \exp(\mathbf{X}_i^T \boldsymbol{\beta}) K_h(X_i - x_0) \mathbf{X}_i \mathbf{X}_i^T.$$

Thus the one-step estimator for $\boldsymbol{\beta}(x_0)$ is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{OS}} = \hat{\boldsymbol{\beta}}_0 - \left[Q''(\hat{\boldsymbol{\beta}}_0)\right]^{-1} Q'(\hat{\boldsymbol{\beta}}_0)$$

and the corresponding estimated covariance matrix for $\hat{\boldsymbol{\beta}}_{\mathrm{OS}}$ is

$$\mathrm{cov}\{\hat{\boldsymbol{\beta}}_{\mathrm{OS}}\} = \left[Q''(\hat{\boldsymbol{\beta}})\right]^{-1} \mathbf{V}(\hat{\boldsymbol{\beta}}) \left[Q''(\hat{\boldsymbol{\beta}})\right]^{-1},$$

where

$$\mathbf{V}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \exp(\mathbf{X}_i^T\boldsymbol{\beta}) K_h^2(X_i - x_0)\mathbf{X}_i\mathbf{X}_i^T.$$

Example 7.5.1. In this example, the covariate X is from an equally spaced design on $[0,\ 10\pi]$, and the conditional distribution of Y given X is a Poisson distribution with mean function

$$\lambda(x) = \exp\{\cos(x)\}.$$

Sample sizes are taken to be 200 and 500. Figures 7.3 and 7.4 show the SiZer maps. The top panels of Figures 7.3 and 7.4 depict the SiZer map based on a maximum local quasi-likelihood with full iterations. The middle ones are the SiZer map based on a one-step local quasi-likelihood estimate with the least-squares estimate as initial estimator, as proposed by Fan and Chen (1999). The bottom ones present the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 7.4.1. From Figures 7.3 and 7.4, it may be concluded that the SiZer map based on a one-step estimator with initial values proposed in Section 7.4.1 is more efficient in distinguishing features than the one with a least-squares estimate as initial values, though their performance are almost the same when the sample size is relatively large. It can be seen from Figures 7.3 and 7.4 that the SiZer map based on a one-step local quasi-likelihood estimate with the proposed initial values performs as well as the one based on a maximum local quasi-likelihood estimate with full iterations. Thus it is recommended to use the SiZer map based on a one-step estimate with the proposed initial values.

7.5.2 Logistic regression

For a Bernoulli distribution, the mean function is the probability function $p(x) = P(Y = 1|X = x)$, the variance function is $p(x)\{1 - p(x)\}$ and the canonical link is logit, i.e. $\mathrm{logit}\{p(x)\} = p(x)/\{1-p(x)\}$. Denote by $Q(\boldsymbol{\beta})$ the local likelihood based on a random sample $\{(X_i, Y_i)\}_{i=1}^n$, then

$$Q'(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i \left\{ Y_i - \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{1 + \exp(\mathbf{X}_i^T\boldsymbol{\beta})} \right\} K_h(X_i - x_0),$$

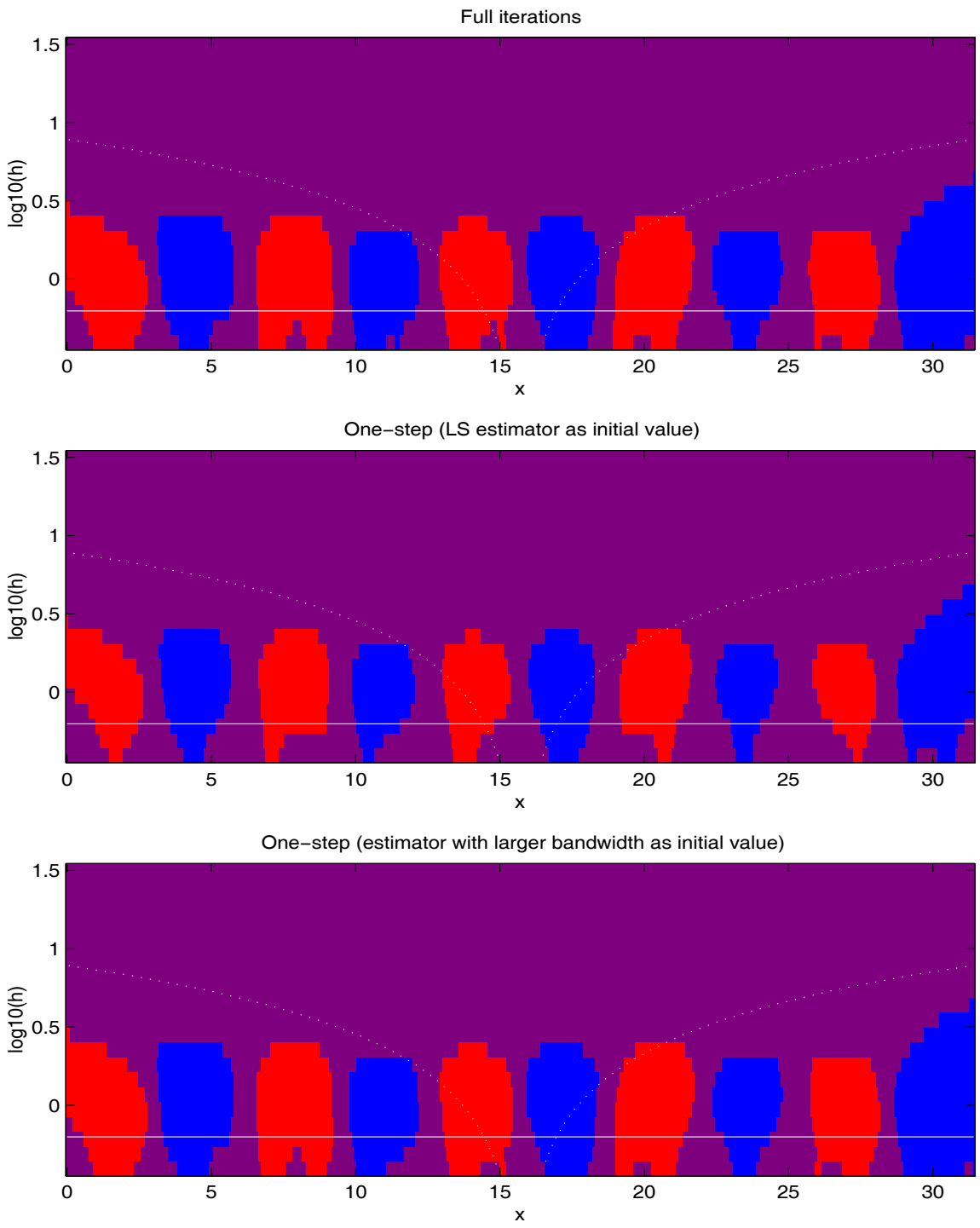


Figure 7.4: *SiZer maps for Poisson regression with sample size* $n = 500$. *The top panel is the SiZer map based on a maximum local quasi-likelihood with full iterations, the middle panel is the SiZer map based on a one-step local quasi-likelihood estimator with a least-square estimate proposed by Fan and Chen (1999) as initial values, and the bottom panel is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 7.4.1.*

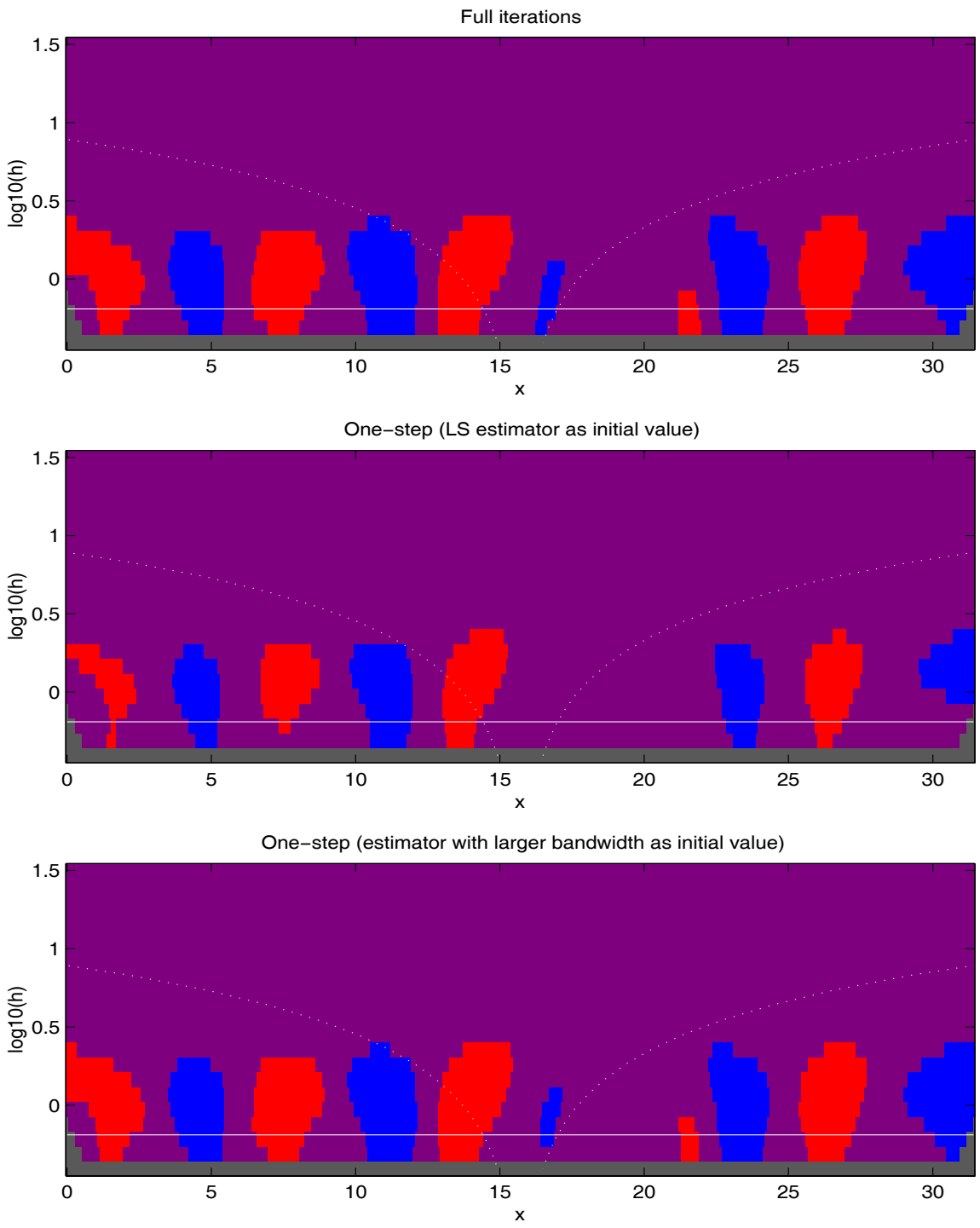


Figure 7.5: *SiZer maps for Poisson regression with sample size $n = 200$. The top panel is the SiZer map based on a maximum local quasi-likelihood with full iterations, the middle panel is the SiZer map based on a one-step local quasi-likelihood estimator with a least-square estimate proposed by Fan and Chen (1999) as initial values, and the bottom panel is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 7.4.1.*

and

$$Q''(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{\{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})\}^2} K_h(X_i - x_0)\mathbf{X}_i\mathbf{X}_i^T.$$

Thus the one-step estimator for $\boldsymbol{\beta}(x_0)$ is given by

$$\hat{\boldsymbol{\beta}}_{\text{OS}} = \hat{\boldsymbol{\beta}}_0 - \left[Q''(\hat{\boldsymbol{\beta}}_0)\right]^{-1} Q'(\hat{\boldsymbol{\beta}}_0)$$

and the corresponding estimated covariance matrix for $\hat{\boldsymbol{\beta}}_{\text{OS}}$ is

$$\text{cov}\{\hat{\boldsymbol{\beta}}_{\text{OS}}\} = \left[Q''(\hat{\boldsymbol{\beta}})\right]^{-1} \mathbf{V}(\hat{\boldsymbol{\beta}}) \left[Q''(\hat{\boldsymbol{\beta}})\right]^{-1},$$

where

$$\mathbf{V}\{\boldsymbol{\beta}\} = \sum_{i=1}^{n} \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{\{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})\}^2} K_h^2(X_i - x_0)\mathbf{X}_i\mathbf{X}_i^T.$$

Example 7.5.2. In this example, the covariate X is equally spaced design on $[0,\ 10\pi]$, and the conditional distribution of Y given X is from Bernoulli with probability function $p(x)$, where

$$\text{logit}\{p(x)\} = \cos(x).$$

In this example, the sample is taken as 500. The SiZer maps are depicted in Figure 7.6. the same conclusion as that in Example 7.5.1 may be drawn from Figure 7.6.

7.5.3 Application

The proposed SiZer map in this section is applied to the environmental data set, analyzed in the last chapter. The association between levels of air pollutant and the number of daily hospital admissions for circulation and respiration problems is of particular interest. As an illustration, we consider how the number of hospital admissions depends on levels of Nitrogen Dioxide NO_2 and how conditional probability of hospital admissions depends on the levels of dust. First the number of hospital admissions is taken as the response variable, and the levels of NO_2 as covariate. Assume that the conditional distribution given the levels of NO_2 is Poisson. The SiZer map is depicted in Figure 7.7. From Figure 7.7, the number of admissions globally increases as the levels of NO_2 increases. However, there are some small wiggles when the smoothing parameters are small. This suggests that the number of admissions also depends on other factors, such as levels of Sulfur dioxide and dust as analyzed in the last Chapter.

The sample mean of the number of admissions is about 250. Figure 7.8 illustrates the SiZer map for the conditional probability that the number of admissions is greater than 250 given

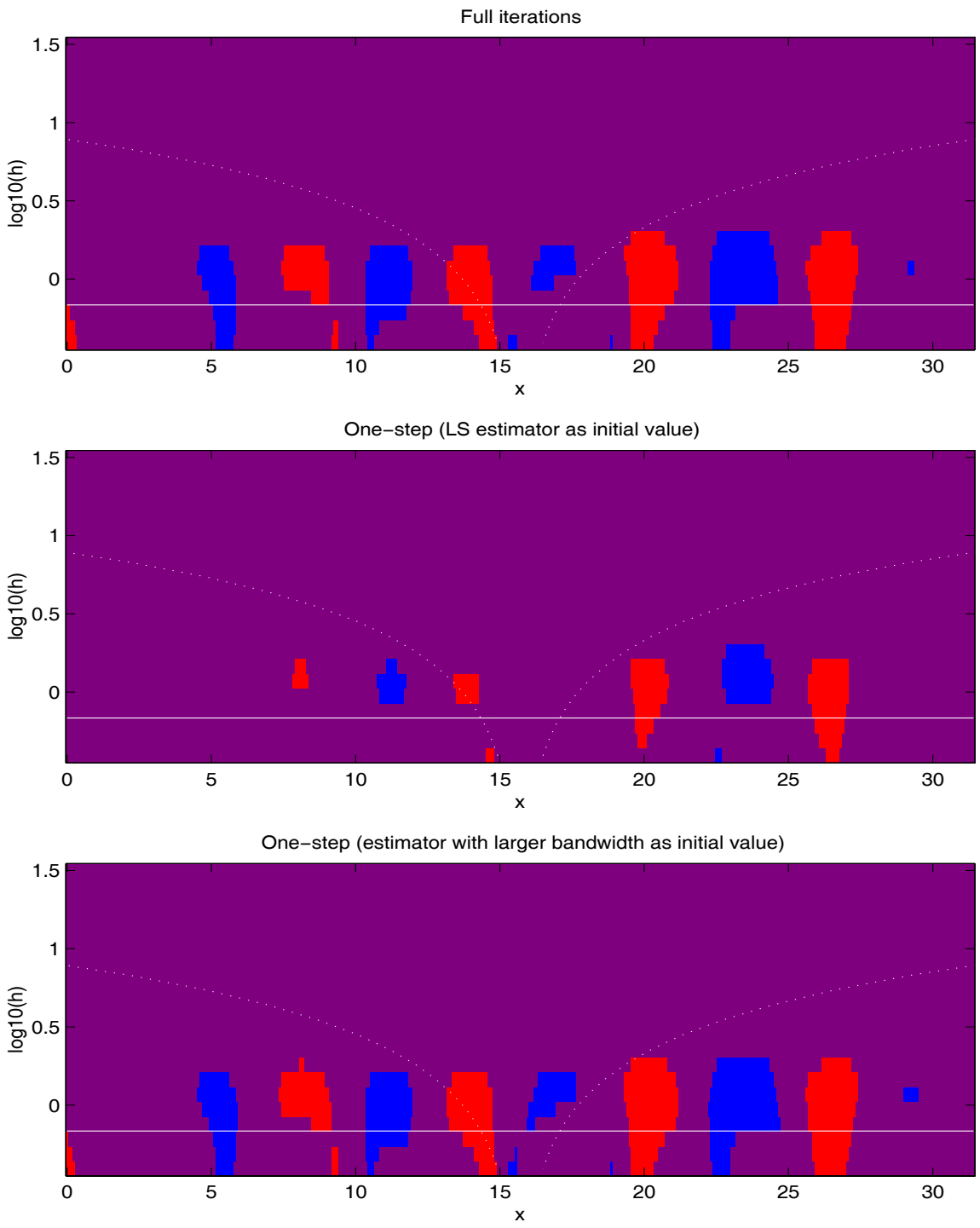


Figure 7.6: *SiZer maps for logistic regression with sample size $n = 200$. The top panel is the SiZer map based on a maximum local quasi-likelihood with full iterations, the middle panel is the SiZer map based on a one-step local quasi-likelihood estimator with a least-square estimate proposed by Fan and Chen (1999) as initial values, and the bottom panel is the SiZer map based on a one-step local quasi-likelihood estimator with initial values proposed in Section 7.4.1.*

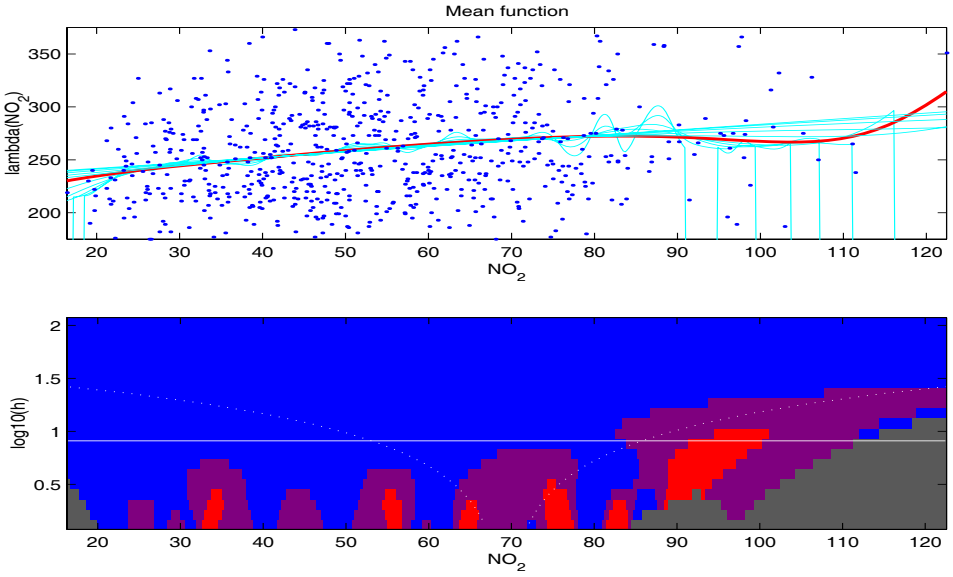


Figure 7.7: *SiZer map for the regression function of the number of admissions. The top panel is the plot of the family smoothing (See Chaudhuri and Marron (1999) for details). The bottom panel depicts the SiZer map for the regression function of the number of admissions given the level of* NO_2.

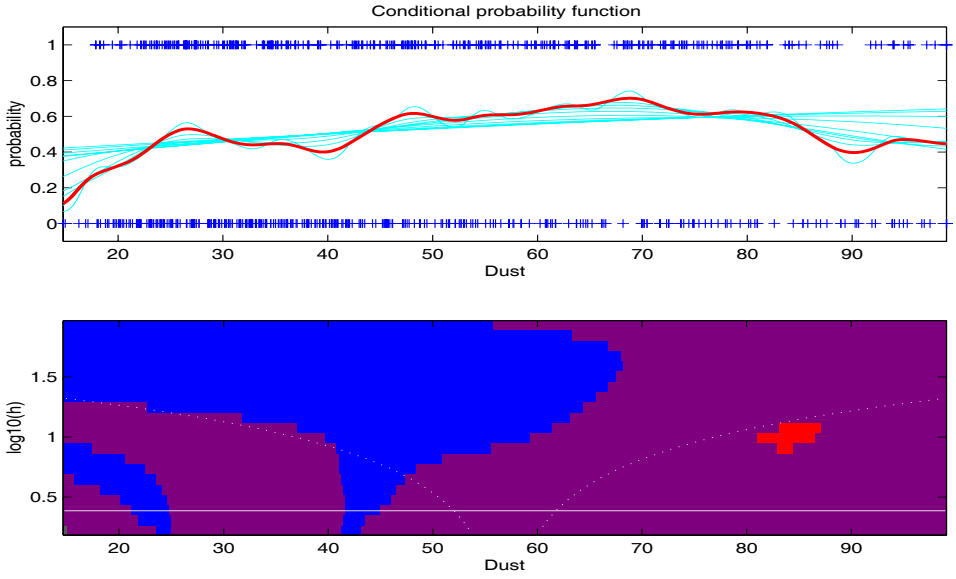


Figure 7.8: *SiZer map for conditional probability of the number of admissions. The top panel is the plot of family smoothing. The bottom panel depicts SiZer map for conditional probability of the number of admissions given the level of dust.*

the level of dust. The conditional probability function globally increases as the level of dust increase from 15 to 65, and then not significantly changes when the level of dust is between 65 to 100, although it seems that it significantly decreases when the level of dust is around 85 and the smoothing parameter is moderate size. Further study is necessary.

7.6 Discussion

In this chapter, the SiZer map proposed by Chaudhuri and Marron (1999) has been extended to the setting of local quasi-likelihood. It has been shown that the SiZer map based on local quasi-likelihood is more efficient, compared with direction application of the proposal of by Chaudhuri and Marron (1999) when quasi-likelihood function are available. An extension of the SiZer map to the setting of local partial likelihood in nonparametric Cox's regression has been tried. However, compared with the SiZer map based on local quasi-likelihood, the SiZer map based on local partial likelihood is very expensive in computation. The reason is that the binned method could not be effectively applied in this context.

Bibliography

[1] Aalen, O. O. (1978). Non-parametric inference for a family of counting processes, *Ann. Statist.*, **6**, 701-726.

[2] Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, **22**, 203–217.

[3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control AC*, **19**, 716–723.

[4] Antoniadis, A. (1999). Wavelets in Statistics: A Review. *Italian Jour. Statist.*, to appear.

[5] Antoniadis, A. and Fan, J. (1999). Regularized wavelet approximations. Submitted .

[6] Bailey, K. R. (1983). The asymptotic joint distribution of regression and survival parameter estimators in the Cox regression model, *Ann. Statist.*, **11**, 39-58.

[7] Bickel, P.J. (1975). One-step Huber estimates in linear models. *J. Amer. Statist. Assoc.*, **70**, 428-433.

[8] Bickel, P.J. (1983). Minimax estimation of a normal mean subject to doing well at a point. In *Recent Advances in Statistics* (M.H. Rizvi, J.S. Rustagi, and D. Siegmund, eds), 511–528. Academic Press, New York.

[9] Breslow, N. E. (1972). Contribution to the discussion on the paper by D. R. Cox, 'Regression and life Table', J. Royal Statist. Soc, B, **34**, 216-217.

[10] de Boor, C. (1978). *A Practical Guide to Splines.* Springer-Verlag, New York.

[11] Bowman, A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis, The Kernel Approach with S-Plus Illustrations.* Oxford Science Publications, Oxford.

[12] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373-384.

[13] Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *Ann. Statist.*, **24**, 2350-2383.

[14] Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves, *J. Amer. Statist. Assoc.*, **93**, 961–976.

[15] Cai, Z., Fan, J. and Yao, Q. (1999). Functional-coefficient regression models for nonlinear time series. To appear in *J. Amer. Statist. Assoc.*

[16] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models, *J. Amer. Statist. Assoc.*, **92**, 477-489.

[17] Carroll, R. J., Ruppert, D. and Welsh, a. H. (1998). Local estimating equations, *J. Amer. Statist. Assoc.*, **93**, 214-227.

[18] Čencov, N.N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Math.*, **3**, 1559–1562.

[19] Chauduhuri, P. and Marron, J.S. (1999). SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.*, **94**, 807–823.

[20] Chen, H. (1988). Convergence rates for parametric components in a partly linear model, *Ann. Statist.*, **16**, 136-146.

[21] Chen, R. and Tsay, R.S. (1993). Functional-coefficient autoregressive models, *J. Amer. Statist. Assoc.*, **88**, 298-308.

[22] Chu, C.K., and Marron, J.S. (1991). Choosing a kernel regression estimator (with discussions). *Statist. Sci.*, **6**, 404–436.

[23] Chui, C.K. (1992). *An Introduction to Wavelets*. Academic Press, Boston.

[24] Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829–836.

[25] Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992). Local regression models, in *Statistical Models in S* (Chambers, J.M. and Hastie, T.J., eds), 309–376, Pacific Grove, California: Wadsworth & Brooks.

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

[27] Cox, D. R. (1974). Partial likelihood. *Biometrika*, **62**, 269-276.

[28] 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. *Numer. Math.*, **31**, 377–403.

[29] Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia.

[30] Donoho, D.L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.*, **22**, 238–270.

[31] Donoho, D.L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harm. Anal.*, **2**, 101–126.

[32] Donoho, D.L. and Johnstone, I.M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

[33] Donoho, D. L. and Johnstone, I. I. (1994b). Minimax risk over ℓ_p balls for ℓ_q error. *probability Theory and Related Fields*, **99**, 277-303.

[34] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, **90**, 1200–1224.

[35] Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 879-921.

[36] Donoho, D. L., Johnson, I.M., Hock, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object (with discussion). *J. Royal Statist. Soc. B*, **54**, 41-81.

[37] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Royal Statist. Soc. B*, **57**, 301–369.

[38] Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

[39] Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, **87**, 998–1004.

[40] Fan, J. (1993). "Local linear regression smoothers and their minimax," *Ann. Statist.*, **21**, 196–216.

[41] Fan, J. (1999). Comments on "Wavelets in statistics: a review" by A. Antoniadis. *J. of Ital. Statist. Assoc.*, To appear.

[42] Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation, *J. Royal Statist. Soc. B*,

[43] Fan, J., Farmen, M. and Gijbels, I. (1998). Local Maximum Likelihood Estimation and Inference, *J. Royal Statist. Soc. B*, **60**, 591-608.

[44] Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.*, **20**, 2008–2036.

[45] Fan, J. and Gijbels, I. (1994). Censored regression: local linear approximations and their applications. *J. Amer. Statist. Assoc.*, **89**, 560–570.

[46] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in olcal polynomial fitting: variable bandwidth and spatial adaptation. *I. Royal. Statist. Soc. B*, **57**, 371-394.

[47] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*, Chapman and Hall, London.

[48] Fan, J., Gijbels, I. and King, M. (1997). Local likelihood and local partial likelihood in hazard regression. *Ann. Statist.*, **25**, 1661-1690.

[49] Fan, J., Hall, P., Martin, M. and Patil, P. (1996). On the local smoothing of nonparametric curve estimators. *J. Amer. Statist. Assoc.*, **91**, 258 – 266.

[50] Fan, J., Hall, P., Martin, M. and Patil, P. (1999). Adaptation to high spatial inhomogeneity based on wavelets and on local linear smoothing. *Statistica Sinica*, **9**, 85 – 102.

[51] Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.*, **90**, 141–150.

[52] Fan, J. and Marron, J. s. (1994). Fast implementations of nonparametric curve estimators, *J. Comput. and Graph. Statist.*, **3**, 35-56.

[53] Fan, J. and Zhang, J. (2000). Functional linear models for longitudinal data, *J. Royal Statist. Soc. B*, to appear.

[54] Fan, J. and Zhang, W. (2000). "Statistical estimation in varying-coefficient models," *Ann. Statist.*, to appear.

[55] Fan, J., Zhang, C. and Zhang, J. (1999). Sieve likelihood ratio statistics and Wilks phenomenon, *Technical Report*, Department of Statistics, UCLA.

[56] Faraggi, D. and Simon, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, **54**, 1475-1485.

[57] Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.

[58] Fu, W. J. (1998). Penalized regression: the bridge versus the LASSO, *J. Comput. Graph. Statist.*, **7**, 397 – 416.

[59] Gao, H. Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.*, **7**, 469–488.

[60] Gao, H. Y. and Bruce, A. G. (1997). WaveShrink with firm Shrinkage. *Statistica Sinica*, **7**, 855 – 874.

[61] Gasser, T., Kneip, a. and Köhler, W., (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.*, **86**, 643-652.

[62] Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. of Statist.*, **11**, 171–185.

[63] González-manteiga, W., Cao, R. and Marron, J. S. (1996). Bootstap selection of the smoothing parameter in nonparametric hazard rate estimation. *J. Amer. Statist. Assoc.*, **91**, 1130-1140.

[64] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.* Chapman and Hall, London.

[65] Hall, P. and Patil, P. (1995a). On wavelet methods for estimating smooth functions. *Bernoulli*, **1**, 41–58.

[66] Hall, P. and Patil, P. (1995b). Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. *Ann. Statist.*, **23**, 905–928.

[67] Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Boston.

[68] Härdle, W., Kerkyacharian, G., Pickard, D., and Tsybakov, A. (1998). *Wavelts, Approximation, and Statistical Applications*. Lecture Notes in Statistics 129, springer-Verlag, New York.

[69] Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests*. Springer, New York.

[70] Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models (with discussion), *J. Royal Statist. Soc. B*, **55**, 757-796.

[71] Hocking, R. R. (1972). Criteria for selection of a subset regression: which one should be used? *Technometrics*, **14**, 967-970.

[72] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data, *Biometrika*, **85**, 809–822.

[73] Huber, P. (1981). *Robust estimation*, Wiley, New York.

[74] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a). A brief survey of bandwidth selection for density esitmation, *J. Amer. Statist. Assoc.*, **83**, 941-953.

[75] Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b). Progress in data-based bandwidth selection ofr kernel density estiamtion, *Computational Statistics*, **11**, 337-381.

[76] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**, 457-481.

[77] Kimeldorf, G.S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.*, **41**, 495–502.

[78] Kimeldorf, G.S. and Wahba, G. (1971). Some results on Tchebysheffian spline functions. *J. Math. Anal. Appl.*, **33**, 82–95.

[79] Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis*. Springer, New York.

[80] Kooperberg, C. and Stone, C.J. (1991). A study of logspline density estimation. *Comp. Statist. & Data Anal.*, **12**, 327–347.

[81] Kooperberg, C., Stone, C.J. and Truong, Y.K. (1995a). Logspline estimation of a possibly mixed spectral distribution. *J. Time Series Anal.*, **16**, 359–388.

[82] Kooperberg, C., Stone, C.J. and Truong, Y.K. (1995b). Rate of convergence for logspline spectral density estimation. *J. Time Series Anal.*, **16**, 389–401.

[83] Lee, E. W., Wei, L. J., and Amato, D. A. (1992). A Cox-type regression analysis for large numbers of small groups of correlated failure time observations, 237-248. In *Survival Analysis: State of the Art*, J. P. Klein and P. Geol, eds. Boston: Kluwer Academic Publishers.

[84] Lehmann, E.L. (1983). *Theory of Point Estimation*. Pacific Grove, California: Wadsworth & Brooks/Cole.

[85] Li, K.-C. (1982). Minimaxity of the method of regularization on stochastic processes. *Ann. Statist.*, **10**, 937–942.

[86] Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.*, **13**, 1352–1377.

[87] Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, B*, **30**, 31-66.

[88] Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

[89] Mack, Y.P. and Müller, H.-G. (1989). Convolution type estimators for nonparametric regression estimation. *Statist. Prob. Lett.* **7**, 229–239.

[90] Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[91] 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**, 278-309.

[92] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. Chapman and Hall, London.

[93] Meyer, Y. (1990). *Ondelettes*. Hermann, Paris.

[94] Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London.

[95] Morris, C. N., Norton, E. C. and Zhou, X. H. (1994). Parametric duration analysis of nursing home usage, 231-248 in *Case Studies in Biometry*, eds by Lange, N. Ryan, L., Billard, L, Brillinger, D. Conquests, L. and Greenhouse, J., Wiley, New York.

[96] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

[97] Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.*, **82**, 231–238.

[98] Nadaraya, E.A. (1964). On estimating regression. *Theory Prob. Appl.*, **9**, 141–142.

[99] Nelson, W. (1972). Theory and Appications of hazard plotting for censored failure data, *Technometrics*, **14**, 945-965.

[100] Nielsen, G. G., Gill, R. D., Andersen, P. K., and Sørensen, T. I. A. A. (1992). A counting process approach to maximum likelihood estimator in frailty models, *Scandinavian J. Statist.*, **19**, 25-43.

[101] Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*, Birkhäuser, Boston.

[102] Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.*, **10**, 177–183.

[103] Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J. Royal Statist. Soc. A*, **135**, 370–384.

[104] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions, *ann. Statist.*, **11**, 453-466.

[105] Rice, J.A. and Rosenblatt, M. (1983). Smoothing splines, regression, derivatives and convolution. *Ann. Statist.*, **11**, 141–156.

[106] Robinson, P.M. (1988). The stochastic difference between econometric and statistics, *Econometrica*, **56**, 531-547.

[107] Ruppert, D., Sheather, s. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. amer. Statist. Assoc.,* **90**, 1257-1270.

[108] Ruppert, D. and Wand, M.P. (1994). Multivariate weighted least squares regression, *Ann. Statist.*, **22**, 1346–1370.

[109] Rutkowski, L. (1982). Orthogonal series estimates of a regression function with applications in system identification. In: *Probability and Statistical Inference* (W. Grossmann *et al.*, eds.), 343–347. North Holland, Amsterdam.

[110] Sargent, D. J. (1998). A general framework for random effects survival anlaysis in the Cox proportional hazards setting. *Biometrics*, **54**, 1486-1497.

[111] Schoenberg, I.J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci.* USA, **52**, 947–950.

[112] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

[113] Seifert, B. and Gasser, Th. (1996). Finite-sample variance of local polynomial: Analysis and solutions, *J. Amer. Statist. Assoc.*, **91**, 267-275.

[114] Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics.* Chapman and Hall, London.

[115] Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.*, **89**, 501–511.

[116] Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selectio method for kernel density estimation. *J. Royal Statist. Soc. B*, **53**, 683-690.

[117] Simonoff, J.S. (1996). *Smoothing Methods in Statistics,* Springer, New York.

[118] Sinha, D. (1998). Posterior likelihood methods for multivariate survival data, *Biometrics*, **54**, 1463-1474.

[119] Slud, E. (1982). Consistency and efficiency of inferences with the partial likelihood, *Biometrika*, **69**, 547-552.

[120] Smith, M. S. and Kohn, R. (1996). Nonparametric regression via Bayesian variable selection. *J. Econometrics*, **75**, 317-343.

[121] Speckman, P. (1988). Kernel smoothing in partial linear models, *J. Royal Statist. Soc. B*, **50**, 413-436.

[122] Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.*, **84**, 276–283.

[123] Stone, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5**, 595–645.

[124] Stone, C.J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, **8**, 1348–1360.

[125] Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040–1053.

[126] Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y.K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.*, **25**, 1371–1470.

[127] Strang, G. (1989). Wavelets and dilation equations: a brief introduction. *SIAM Review*, **31**, 614–627.

[128] Strang, G. (1993). Wavelet transforms versus fourier transforms. *Bulletin Amer. Math. Soc.*, **28**, 288–305.

[129] Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. Royal Statist. Soc. B*, **58**, 267-288.

[130] Tibshirani, R. J. (1997). The LASSO method for variable selection in the Cox model. *Statist. Med.*, **16**, 385-395.

[131] Tibshirani, R. and Hastie, T.J. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.*, **82**, 559–567.

[132] Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *J. Royal Statist. Soc. B*, **61**, 529–546.

[133] Tsiatis, A. (1981). A large sample study of Cox's regression model, *Ann. Statist.* **9**, 93-108.

[134] Utreras, F.D. (1980). Sur le choix du parametre d'adjustement dans le lissage par fonctions spline. *Numer. Math.*, **34**, 15–28.

[135] Vadakovic, R. (1999). *Statistical Modeling by Wavelets*, Wiley, New York.

[136] Wahba, G. (1975). Smoothing noisy data with spline functions. *Numer. Math.*, **24**, 383-393.

[137] Wahba, G. (1977). A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (P.R. Krisnaiah, ed.), 507–523. North Holland, Amsterdam.

[138] Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

[139] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*, London: Chapman and Hall.

[140] Watson, G.S. (1964). Smooth regression analysis. *Sankhyā* Ser. A, **26**, 359–372.

[141] Wong, W. H. (1986). Theory of partial likelihood, *Ann. Statist.*, **14**, 88-123.