PENALIZED EMPIRICAL LIKELIHOOD BASED VARIABLE SELECTION

THARSHANNA NADARAJAH

PENALIZED EMPIRICAL LIKELIHOOD BASED VARIABLE SELECTION

by

©Tharshanna Nadarajah

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirement for the Degree of

Master of Science in Statistics

Department of Mathematics and Statistics

Memorial University of Newfoundland

St. John's Newfoundland, Canada July 2011

Abstract

Variable selection is an important topic in high-dimensional statistical modeling, especially in generalized linear models. Several variable selection procedures have been developed in the literature, including the sequential approach, prediction-error approach, and information-theoretic approach. All of these are computationally expensive. A new method based on penalized likelihood has been lauded for its computational efficiency and stability. In this approach the variable selection and the estimation of the coefficients are carried out simultaneously. The parametric likelihood is a crucial component, but in many situations a well-defined parametric likelihood is not easy to construct. To overcome this problem, Variyath (2006) proposed a penalized-empirical-likelihood (PEL) based variable selection where empirical likelihood is constructed based on a set of estimating equations. We investigate the asymptotic properties of the new method, and develop an algorithm for estimating the parameters. Our simulation studies show that when a parametric model is available,

PEL-based variable selection gives results similar to those achieved by parametric-likelihood variable selection. The former method outperforms the latter when the parametric model is misspecified. We extend our approach to variable selection in Cox's proportional hazard model.

Acknowledgements

I would like to express my appreciation to my supervisor Dr. Asokan Mulayath Variyath, for giving me the opportunity to work with him. His continuous guidance, support and patience over the last two years have been invaluable.

I gratefully acknowledge the financial support provided by Memorial University of Newfoundland's School of Graduate Studies, the Department of Mathematics & Statistics, and my supervisor in the form of graduate fellowships and teaching assistantships. I would like to thank all the faculty in our department for their support. It would not have been possible for me to complete the requirements of the graduate program without their guidance.

I express my profound appreciation to my parents, brothers, wife Premni, and beautiful son Kavevarman for their encouragement, understanding, and patience even during the difficult times.

Last but not least, it is my great pleasure to thank the friends and well-wishers

who directly or indirectly encouraged me in the Master's program and contributed to this dissertation.

Finally, this thesis is dedicated to my parents and the teachers who have supported me ever since the beginning of my studies.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of Variable Selection

Variable selection is an important topic in statistical modeling, especially in generalized linear models (GLM). In practice, a large number of covariates, $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p)$, are believed to have an influence on the response variable $\boldsymbol{y}$ of interest. However, some covariates have no influence or a weak influence, and a regression model that includes all the covariates is not advisable. Excluding the unimportant covariates results in a simpler model with better interpretive and predictive value.

The problem of identifying a submodel that adequately models the response is generally referred to as the variable selection problem. Statistically speaking, variable

selection is a way to reduce the complexity of the model, in some cases by accepting a small amount of bias to improve the precision. The main advantages of selecting a subset of the variables instead of the entire set are:

- ◇ The interpretation of a large model can be difficult.
- ◇ The prediction accuracy may be improved by dropping redundant and irrelevant variables.
- ◇ Knowing which variables are significant gives insight into the nature of the prediction problem and allows a better understanding of the final model.
- ◇ It is cheaper to measure a reduced set of variables.

For example, consider the doctor-visit data from the Australian health survey of 1977–78, which is discussed in detail by Cameron and Trivedi (1998). The data set consists of a response variable (the number of doctor visits in the previous two weeks by an adult) and twelve covariates, including health indicators and general factors, which are listed in Table 1.1. Our goal is to model the relationship between the response and the covariates. The model with all covariates is not interesting since it is difficult to interpret and will have poor prediction precision. We aim to find a simpler model that gives a reasonable description of the data-generating mechanism. The initial analysis of and variable selection for this data set are discussed in Chapter 5. In

the next subsection we will discuss commonly used regression models and estimation procedures where variable selection is considered important.

Variables	Description
y-Dvisits	Number of doctor visits in previous two weeks
X_1-Sex	1 if female, 0 if male
X_2-Age	Age in years divided by 100
X_3-Agesq	Age squared
X_4-Income	Annual income in Australian dollars divided by 1000
X_5-Levyplus	1 if covered by private health insurance; 0 otherwise
X_6-Freepoor	1 if covered by government because low income, recent immigrant, unemployed; 0 otherwise
X_7-Freerepa	1 if covered free by government because elderly, disability pension, invalid veteran, or family of deceased veteran; 0 otherwise
X_8-Illness	Number of illnesses in previous 2 weeks, with 5 or more coded as 5
X_9-Actdays	Number of days of reduced activity in previous 2 weeks due to illness or injury
X_{10}-Hscore	General health questionnaire score using Goldberg's method; high score indicates bad health
X_{11}-Chcond1	1 if chronic condition(s) but not limited in activity; 0 otherwise
X_{12}-Chcond2	1 if chronic condition(s) and limited in activity; 0 otherwise

Table 1.1: Response and covariates of doctor-visit data

1.1.1 Linear Models

Linear models have been the mainstay of statistics for thirty years and remain one of our most commonly used statistical tools. In linear models, the data are modeled using linear functions of the covariates, and the unknown parameters are estimated from the data. For a given data set $\{y_i; x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n units/subjects, a linear

regression model assumes that the relationship between the response variable y_i and the p dimensional regressors $\boldsymbol{X_i}$ is linear. Thus, the model has the form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

where ϵ is the error term, $\boldsymbol{X}$ is an $n \times p$ matrix of covariate values, and $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated. A violation of the linearity assumption between the response and the explanatory variables or the distributional assumption of the random error may increase the model variation. The method of least squares is the most popular method for estimating the regression parameters. This approach minimizes the residual sum of squares,

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

In matrix form, the residual sum of squares can be written

$$\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{1.2}$$

Hence, the ordinary least-squares estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{1.3}$$

and the fitted values at the training inputs are

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

If we assume that $\boldsymbol{\epsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$, then the likelihood function of $\boldsymbol{y}$ can be written

$$\mathbf{L}(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\sum_{i=1}^{p} (y_i - \hat{y}_i)^2 \right\}.$$

Let $\ell(\boldsymbol{\beta}, \sigma^2) = \log \mathbf{L}(\boldsymbol{\beta}, \sigma^2)$, then the partial derivative of $\ell(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})$$

and setting $\dfrac{\partial \ell(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} = 0$ gives an estimate of $\boldsymbol{\beta}$, which is same as the least-squares estimate of $\boldsymbol{\beta}$.

1.1.2 Generalized Linear Model (GLM)

Generalized linear models are defined by Nelder and Wedderburn (1972). GLMs include linear regression models, logistic and probit models for categorical responses, and log-linear models. For all these models, a linear relationship is assumed between the response variable $\boldsymbol{y}$ and covariates $\boldsymbol{X}$ through some link function. The conditional expectation of $\boldsymbol{y}$ given $\boldsymbol{X}$ is specified as

$$\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{X}) = g(\boldsymbol{X}\boldsymbol{\beta}), \tag{1.4}$$

where $g(*)$ is a known link function and $\boldsymbol{\beta}$ is the vector of regression parameters. A GLM includes a random component specifying the conditional distribution of the

response variable $\boldsymbol{y}$ given the explanatory variable. Traditionally, the random component is a member of an exponential-family distribution such as the Gaussian, binomial, Poisson, gamma, or inverse-Gaussian. The estimation proceeds by defining a measure of goodness-of-fit between the observed data and the fitted values generated by the model. The parameter estimates are the values that minimize the goodness-of-fit criterion. We primarily estimate the parameters by maximizing the likelihood for the observed data. The log-likelihood based on a set of independent observations $y_1, y_2, \ldots, y_n$ is

$$\ell(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\beta}).$$

The goodness-of-fit criterion is .

$$D(\boldsymbol{y}; \boldsymbol{\mu}) = 2\ell(\boldsymbol{y}; \boldsymbol{y}) - 2\ell(\boldsymbol{\mu}; \boldsymbol{y});$$

it is called the *scaled deviance*. Note that $\ell(\boldsymbol{y}; \boldsymbol{y})$ is the maximum likelihood for an exact fit in which the fitted values are equal to the observed data, and it does not depend on the parameters. Maximizing $\ell(\boldsymbol{\mu}; \boldsymbol{y})$ is equivalent to minimizing $D(\boldsymbol{y}; \boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$, subject to the constraints imposed by the model.

1.1.3 Quasi-Likelihood (QL)

When there is insufficient information about the data for us to specify a parametric model, quasi-likelihood is often used. In this situation we can develop the statistical analysis based on approximations to the likelihood, and we concentrate on cases where the observations are independent. Suppose we have a vector of independent responses, $\boldsymbol{y}$, with mean $\boldsymbol{\mu}$ and covariance diagonal matrix $\sigma^2 \mathrm{V}(\boldsymbol{\mu})$. We assume that $\boldsymbol{\mu}$ is a function of covariates and some regression parameters $\boldsymbol{\beta}$. To construct the quasi-likelihood, we start by looking at a single component $\boldsymbol{y}$ of y. Under the above conditions, the function

$$U = u(\mu; y) = \frac{y - \mu}{\sigma^2 \mathrm{V}(\mu)}$$

has the following properties:

$$E(U) = 0,\ \mathrm{V}(U) = \frac{1}{\sigma^2 \mathrm{V}(\mu)},\ \text{and}\ -E\left(\frac{\partial U}{\partial \mu}\right) = \frac{1}{\sigma^2 \mathrm{V}(\mu)}.$$

Most of the first-order asymptotic theory concerned with the likelihood is based on these properties. It is therefore not surprising that

$$\mathbf{Q}(\mu; y) = \int_y^{\mu} \frac{y - t}{\sigma^2 \mathrm{V}(t)} dt$$

behaves like a log-likelihood function for μ; this is called the quasi-likelihood. The quasi-likelihood for complete data is

$$\mathbf{Q}(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{n} \mathbf{Q}(\mu_i; y_i).$$

The quasi-deviance function for a single observation can be written

$$D(\mu; y) = -2\sigma^2 \mathbf{Q}(\mu; y) = 2 \int_{\mu}^{y} \frac{y - t}{\mathrm{V}(t)} dt.$$

The quasi-likelihood estimating equations for the regression parameters $\boldsymbol{\beta}$ are obtained by differentiating $\mathbf{Q}(\boldsymbol{\mu}; \boldsymbol{y})$. They can be written in the form $U(\hat{\boldsymbol{\beta}}) = 0$, where

$$U(\boldsymbol{\beta}) = \frac{\mathbf{D}^T \mathrm{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})}{\sigma^2}$$

is called the quasi-score function and $\mathbf{D}$ is the derivative of $\boldsymbol{\mu}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The Newton-Raphson method is widely used to estimate the parameters.

1.2 Variable Selection Methods

The main objective of variable selection methods is to identify a simpler adequate model that is easier to interpret than the full model. In linear models, the submodel relates the response variable $\boldsymbol{y}$ to a subset of components of $\boldsymbol{X}$ in the form

$$\boldsymbol{y} = \boldsymbol{X}(s)\boldsymbol{\beta}(s) + \boldsymbol{\epsilon}$$

where $\boldsymbol{X}(s)$ is a subset of the components of $\boldsymbol{X}$, $\boldsymbol{\beta}(s)$ is a vector of the corresponding regression parameters, and $s \subseteq (1, 2, \ldots, p)$. The variable selection problem

is to find the best subset s such that the submodel is optimal according to some criterion that gives a good description of the data-generating mechanism. Several methods have been developed in the literature for the identification of the best submodel. These methods can be broadly classified into four categories: sequential approaches, prediction-error approaches, information-theoretic approaches, and penalized-likelihood approaches. In the next section we will discuss existing variable selection procedures and their advantages and disadvantages.

1.2.1 Sequential Approaches

The sequential approaches were developed in the early 1960s when computing resources were limited. In these approaches, only some of the possible submodels are evaluated to identify the best model. In the forward-selection approach, we start with an intercept model and add the variables one at a time. At each step, each variable that is not already in the model is tested for inclusion, and the most significant variable is added to the model. This process continues until none of the remaining variables are significant when added to the model or there are no more variables. Because of the complexity that arises from the nature of this procedure, it is essentially impossible to control the error rate.

Forward selection has drawbacks, including the fact that addition of a new variable

may change the significance of one or more variables already included in the model. An alternative approach is backward elimination. In this approach, we start a model with all the variables of interest. Then the least important variable is dropped, provided it is not significant. We continue this process by successively re-fitting reduced models and applying the same rule until all the variables remaining in the model are statistically significant. Backward elimination also has drawbacks. Sometimes variables that are dropped would be significant in the final reduced model. This suggests that a compromise between forward selection and backward elimination should be considered.

Efroymson (1960) proposed a stepwise-regression approach that is a combination of the above two approaches. This method uses forward selection, but after the addition of each variable, backward elimination is applied to potentially remove variables already in the model. Stepwise regression does not guarantee to find an optimal submodel. The sequential approaches are computationally less demanding than the other methods.

1.2.2 Prediction-Error Approach

Another approach to variable selection is to choose the submodel with the best ability to predict a future response. Methods using the prediction-error approach, such as

cross-validation and bootstrap, are computationally intensive. Cross-validation has been well studied as a basis for model selection by Stone (1974). In cross-validation, we compute the prediction error of all submodels. We split the data into K parts of roughly equal sizes and estimate the prediction error for one part of the data based on the fitted submodel using the remaining $(K-1)$ parts. We then combine all K estimates of the prediction error for each submodel. The submodel with the minimum prediction error is selected.

Let $k : \{1, 2, \ldots, n\} \mapsto \{1, 2, \ldots, K\}$ be an indexing function that indicates the partition to which each observation is allocated by the randomization. The case $K = n$ is known as leave-one-out cross-validation. In this case the cross-validation estimators are approximately unbiased for the true prediction error, but they can have a high variance and the computational burden is also high. In general, five- or ten-fold cross-validation is recommended (see Breiman and Spector, 1992; Kohavi, 1995).

Bickel and Freedman (1982) suggested that conditional bootstrap be used for variable selection. The bootstrap is a general tool for assessing statistical accuracy. Suppose we wish to fit a model to a set of training data. The basic idea is to randomly draw data sets with replacement from the training data, each of the same size as the original training set. This procedure repeated a large number of times. Then we refit

the model to each of the bootstrap sample sets and examine the behavior of the fits.

These methods are computer-intensive and tend to be impractical if we have to fit more than 15–20 models or if the sample size is large. However, cross-validation offers an interesting alternative for model selection. In some situations the prediction error is not well defined (for example, in generalized linear models) and therefore these methods are not applicable.

1.2.3 Information-Theoretic Approach

In this section, we briefly introduce the most commonly used information-theoretic model selection approaches: the Akaike information criterion (AIC) and Bayesian information criterion (BIC). These methods are applicable when a well-defined parametric model is available. We will also discuss nonparametric versions of AIC and BIC.

Akaike Information Criterion (AIC)

Kullback and Leibler (1951) introduced the Kullback-Leibler (K-L) "distance" or "information" between two models. Let f and g be continuous distribution functions, then the K-L information between models f and g is defined to be

$$I(f, g) = \int f(x) \log \left[\frac{f(x)}{g(x|\theta)} \right] dx.$$

The notation $I(f, g)$ denotes the distance from g to f. However, the K-L distance can not be computed without full knowledge of both f and the parameter θ for each candidate model $g_i(x|\theta)$. Akaike (1973, 1974) found a simple relationship between the K-L distance and Fisher's maximized log-likelihood function. Akaike also found a rigorous way to estimate the K-L information, based on the empirical log-likelihood function at its maximum point. We represent the full model with p parameters as

$$\text{model}(p): \ f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}_p), \ \boldsymbol{\beta}_p = (\beta_1, \beta_2, \ldots \beta_s, \beta_{s+1}, \ldots, \beta_p)^T.$$

Akaike formulates the problem of statistical model identification as the selection of a submodel $f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}_s)$, where the particular restricted model is defined by the constraints $\beta_{s+1} = \beta_{s+2} = \ldots = \beta_p = 0$, so that

$$\text{model}(s): \ f(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}_s), \ \boldsymbol{\beta}_s = (\beta_1, \beta_2, \ldots \beta_s, 0, \ldots, 0)^T$$

where s is the number of parameters and $\boldsymbol{\beta}_s$ is a subspace of $\mathbb{R}^p$. Let $\hat{\boldsymbol{\beta}}_s$ be the maximum likelihood estimate under model(s), then the log-likelihood function is given by

$$\ell(\hat{\boldsymbol{\beta}}_s) = \sum_{i=1}^{n} \log \Big\{ f(y_i, X_i, \hat{\boldsymbol{\beta}}_s) \Big\}.$$

The Akaike information criteria of submodel s is defined to be

$$\text{AIC}(s) = -2\ell(\hat{\boldsymbol{\beta}}_s) + 2k$$

where k is the cardinality of s. Under this criterion we choose the model with the minimum AIC value.

Bayesian Information Criterion (BIC)

Schwarz (1978) suggested using a Bayesian approach to the model selection problem. This method results in a criterion that is similar to AIC. It is based on the penalized log-likelihood function evaluated at the maximum likelihood estimate for the model. The penalty term in the BIC obtained by Schwarz (1978) is the AIC penalty term k multiplied by $\frac{1}{2}\log(n)$, where n is the sample size. Similarly to AIC, the BIC of a submodel is defined to be

$$\text{BIC}(s) = -2\ell(\hat{\boldsymbol{\beta}}_s) + k\log(n).$$

The submodel with the minimum BIC value is selected. It has been observed that minimizing AIC does not produce asymptotically consistent estimates of the correct model. In contrast, BIC is consistent.

Mallow's C_k Criterion

Mallow's C_k is a technique for model selection in regression proposed by Mallows (1973, 1995). The C_k statistic is a criterion to assess the fit when models with

different numbers of parameters are being compared. The Mallows criterion for a submodel is

$$C_k(s) = \frac{RSS(s)}{\hat{\sigma}^2} - n + 2k$$

where $RSS(s)$ is the residual sum of squares and k is the cardinality of s. Usually C_k is plotted against k for the collection of subset models of various sizes under consideration. Acceptable models (minimizing the total bias of the predicted values) are those for which C_k approaches the value k.

In summary, the information-theoretic approaches are based on strong parametric model assumptions. In GLMs and QL, the model is frequently specified by a set of estimating equations and we may not have fully specified parametric assumptions. Hence, these methods can not be used directly. One solution is to use nonparametric likelihood based on the available information. Another limitation of the information-theoretic approach is the computational burden of fitting all possible submodels. In the next section, we discuss the empirical-likelihood-based information-theoretic approach for variable selection proposed by Variyath, Chen, and Abraham (2010).

Empirical-Likelihood-Based Information-Theoretic Approach

Variyath, Chen, and Abraham (2010) developed an information-theoretic approach to variable selection based on a nonparametric likelihood, for use when a well-defined

parametric model is not available. They replaced the parametric likelihood by the empirical likelihood and investigated the use of empirical-likelihood-based AIC and BIC. The empirical-likelihood-based AIC is defined to be

$$\text{EAIC}(s) = W(\hat{\boldsymbol{\beta}}_s) + 2k,$$

where $W(\hat{\boldsymbol{\beta}}_s) = 2\ell_{EL}(\hat{\boldsymbol{\beta}}_s)$ is the empirical-likelihood ratio function for the submodel. Similarly, the empirical-likelihood-based BIC is defined to be

$$\text{EBIC}(s) = W(\hat{\boldsymbol{\beta}}_s) + k\log(n).$$

The best model is identified as the model with the minimum value of EAIC (or EBIC) over all possible submodels. More details of the empirical likelihood are given in Chapter 3. Variyath, Chen, and Abraham (2010) show that the empirical and parametric likelihood-based AIC and BIC have first-order asymptotic properties. Their simulation studies show that when a parametric likelihood exists, the two methods have similar performance. The empirical-likelihood-based approach is superior when the parametric model is misspecified.

In the information-theoretic approach a complete evaluation of all the submodels is necessary. As the number of covariates increases, the computational burden becomes more severe. To avoid the evaluation of all the submodels, a new penalized-likelihood variable selection approach has recently been developed.

1.2.4 Penalized-Likelihood Approach

The idea of penalization is very useful in statistical modeling particularly in high dimensional variable selection. Most traditional variable selection procedures such as AIC, Mallow's C_k, and BIC use a fixed penalty based on the size of the model. However, all these procedures use either stepwise or subset-selection procedures to select the variables. These selection procedures make the procedures computationally intensive and unstable. To overcome the inefficiencies of traditional variable selection procedures, Fan and Li (2001) proposed a unified approach via nonconcave penalized least squares. This method automatically and simultaneously selects variables and estimates their coefficients. The least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani (1996, 1997) is another variant of the penalized-likelihood approach. Fan and Li (2001) applied the penalized-likelihood approach to linear regression, robust linear regression, and generalized linear models. They show that the proposed penalized-likelihood estimator with the smoothly clipped absolute deviation (SCAD) penalty function (defined in Chapter 2) outperforms all the subset and information-theoretic variable selection procedures in terms of computational cost and stability. The SCAD improves the LASSO by reducing the estimation bias. Furthermore, they show that the SCAD possesses oracle properties with a proper choice of the tuning parameters. The true regression coefficients that are zero are

automatically shrunk to zero, and the remaining coefficients are simultaneously estimated. Hence, the SCAD and its properties are ideal procedures for variable selection, at least from a theoretical point of view. This encourages us to investigate SCAD properties in nonparametric-likelihood setting.

1.2.5 Motivation for New Approach

Several methods have been developed to select the best submodel. The sequential approaches are computationally less demanding as the number of covariates increases, but the identification of the optimal model is not guaranteed. The simplest and most widely used variable selection method is cross-validation. In some situations the prediction error is not well defined, for example in generalized linear models, which limits the application of this technique. Information-theoretic variable selection methods such as AIC and BIC are based on the parametric likelihood. These two criteria can not be applied without full knowledge of the parametric model. If the model is not well defined, we can use empirical-likelihood-based AIC and BIC. In some situations, the number of possible submodels is large, and the computational cost becomes substantial if all the submodels must be evaluated. Methods based on penalized likelihood such as LASSO and SCAD have superior computational efficiency and stability. SCAD improves on LASSO by reducing the estimation bias and it

satisfies the oracle properties. The parametric likelihood is a crucial component of these methods. As discussed earlier, the parametric model is not well defined in many cases, limiting the application of the methods. We investigate the properties of SCAD in a nonparametric setting, where instead of the parametric likelihood, we use the empirical likelihood based on a set of estimating equations.

1.3 Proposed Approach to Variable Selection

Likelihood methods play a major role in statistical analysis. They can be used to find efficient estimators and are flexible. Likelihood methods can reduce or eliminate the problems arising when the data are incompletely observed, distorted, or sampled with a bias. They can be used to pool information from different data sources. One problem with parametric likelihood inference is the risk of model mis-specification. Such mis-specification can cause likelihood-based estimates to be inefficient. To avoid the risk of model mis-specification, a nonparametric method can be used. Instead of parametric likelihood, we use nonparametric empirical likelihood in the penalized-likelihood variable selection approach.

1.3.1 Empirical Likelihood (EL)

Owen (1988) introduced the empirical likelihood. Empirical likelihood is a nonparametric method of statistical inference. It allows us to use likelihood methods without assuming that the data come from a known distribution. The empirical likelihood method combines the reliability of nonparametric methods with the flexibility and effectiveness of the likelihood approach.

Let $y_1, y_2, \ldots, y_n$ be a random sample from a cumulative distribution function $F(*)$. Let

$$p_i = P(y = y_i) = F(y_i) - F(y_i-)$$

be the probability mass assigned to y_i. The empirical likelihood function defined by Owen (1988) is

$$L(\boldsymbol{F}) = \prod_{i=1}^{n} p_i.$$

Maximizing

$$\ell(\boldsymbol{F}) = \log\{L(\boldsymbol{p})\} = \sum_{i=1}^{n} \log(p_i)$$

subject to $p_i > 0$ and $\sum_{i=1}^{n} p_i = 1$ leads to $\hat{p}_i = \frac{1}{n}$. The maximum empirical likelihood (MEL) estimator of $F_n(y)$ is given by

$$\hat{F}(y) = \sum_{i=1}^{n} \hat{p}_i I(y_i \leq y) = F_n(y),$$

where $I(*)$ is the indicator function. The empirical distribution function based on a random sample is

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \leq y).$$

Statistical inference on the parameters can be based on the profile empirical likelihood. For example, if we are interested in inference on the mean, say μ, we define the profile empirical log-likelihood for μ to be

$$\ell(\mu) = \sup\left\{\sum_{i=1}^{n}\log(p_i) : \ p_i > 0, i = 1, 2, \ldots, n; \ \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i(y_i - \mu) = 0\right\}.$$

Owen (1988, 1990, 2001) proved that the empirical likelihood ratio function has an asymptotic χ^2 distribution when $\mu = \mu_0$, the true value. This result is useful for inference on the parameters, such as testing hypotheses and constructing a confidence region for μ. Note that there is no need to estimate a scale parameter in the construction of the confidence interval, and the confidence regions are not necessarily symmetric because of the data-driven approach. Because of these properties, the EL method has become popular in the statistical literature and has been extended to linear regression models (Owen, 1991; Chen, 1993, 1994), general estimating equations (Qin and Lawless, 1994), survival analysis (Thomas and Grunkemeier, 1975; Li, 1995; Murphy, 1995), survey sampling (Chen and Qin, 1993; Chen, Sitter, and Wu, 2002) and time series (Monti, 1997).

1.3.2 Penalized Empirical Likelihood (PEL)

As discussed earlier, penalized-likelihood-based variable selection can be applied only when we have a well-defined parametric model. When we are not sure about the parametric model, but the parameters can be estimated by a set of estimating equations, we can use an EL based on a set of estimating equations. So we propose to replace the parametric likelihood by the empirical likelihood to define a nonparametric version of the penalized likelihood method. We discuss the asymptotic properties of the regression estimates, and we develop an algorithm for estimating the parameters. Our simulation studies show that when a parametric model is available, PEL-based variable selection gives results similar to those achieved by parametric-likelihood variable selection. The former method outperforms the latter when the parametric model is misspecified. We extend our approach to Cox's proportional hazards model. We also apply our method to an Australian health survey and a lung-cancer data set.

1.4 Outline of the Thesis

The main objective of this thesis is to make a contribution to variable selection. We mainly focus on penalized-empirical-likelihood variable selection. In Chapter 2 we briefly discuss variable selection via the nonconcave penalized likelihood proposed

by Fan and Li (2001). In Chapter 3, we introduce the empirical likelihood and its characteristics. We describe our penalized-empirical-likelihood variable selection and discuss its asymptotic properties. The algorithm is given in Chapter 4. In Chapter 5 we provide simulation studies to compare the performance of empirical-likelihood variable selection with penalized-parametric-likelihood SCAD, in the context of linear regression, Poisson regression, and logistic regression. We also apply our method to the Australian health survey. In Chapter 6, we discuss the implementation of PEL in Cox's proportional hazard model. Our concluding remarks are given in Chapter 7.

Chapter 2

Variable Selection via Nonconcave Penalized Likelihood

A new class of variable selection methods based on a nonconcave penalized-likelihood approach was proposed by Fan and Li (2001) and Tibshirani (1996). These methods are superior to traditional methods because of their computational efficiency and stability. The variable selection and the estimation of the regression parameters are carried out simultaneously. That is, insignificant variables are removed by estimating their regression parameters as zero. These methods work reasonably well in high-dimensional problems. In this chapter, we will introduce the penalized-likelihood variable selection proposed by Fan and Li (2001) in the context of a linear model.

Consider a linear model of the form

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i, \ i = 1, 2, \ldots, n,$$

where $\boldsymbol{X}_i \in \mathcal{R}^p$ is a vector of covariates and $\boldsymbol{\beta} \in \mathcal{R}^p$ a vector of parameters. We assume that the collected data $\{(\boldsymbol{X}_i, y_i)\}$ are independent samples and $y_i|\boldsymbol{X}_i$ has density $f(y_i; \boldsymbol{X}_i\boldsymbol{\beta})$. A general form of the penalized likelihood proposed by Fan and Li (2001) is defined by

$$\sum_{i=1}^{n} \ell(y_i; \boldsymbol{X}_i\boldsymbol{\beta}) - n\sum_{j=1}^{p} p_\delta(|\beta_j|) \tag{2.1}$$

where $\ell(y_i; \boldsymbol{X}_i\boldsymbol{\beta})$ is the conditional log-likelihood of $y_i|\boldsymbol{X}_i$, $p_\delta(*)$ is a penalty function, and δ is the tuning parameter.

In linear regression models, if the columns of the design matrix $\boldsymbol{X}$ are orthonormal then it is easy to show that the best-subset selection method and the stepwise elimination method are equivalent to penalized least-squares estimations with the HARD thresholding penalty proposed by Fan (1997) and Antoniadis (1997). This penalty is defined to be

$$p_\delta(|\theta|) = \delta^2 - (|\theta| - \delta)^2 I(|\theta| < \delta).$$

For a large value of $|\theta|$, the HARD thresholding penalty does not overpenalize. The LASSO penalty function is the L_1-penalty, $p_\delta(|\theta|) = \delta|\theta|$, proposed by Donoho and

Johnstone (1994) in the wavelet setting and extended by Tibshirani (1996) to general likelihood settings. The penalty function used in ridge regression is the L_2 penalty, $p_\delta(|\theta|) = \delta|\theta|^2$. According to Fan and Li (2001), a good penalty function should result in an estimator with the following three oracle properties:

1. Unbiasedness: To avoid unnecessary modeling bias, the estimator is nearly unbiased when the true unknown parameter is large.

2. Sparsity: This is a thresholding rule that automatically sets small estimated coefficients to zero to reduce the model complexity.

3. Continuity: This property eliminates unnecessary variation in the model prediction.

However, the penalty functions L_1, L_2, and HARD do not satisfy all three conditions. A simple penalty function satisfying all three is the SCAD penalty proposed by Fan (1997). Its first derivative is

$$p'_\delta(\theta) = \delta \left\{ \mathrm{I}(\theta \leq \delta) + \frac{(a\delta - \theta)_+}{(a-1)\delta} \mathrm{I}(\theta > \delta) \right\} \quad \text{for some } a > 2 \text{ and } \theta > 0. \tag{2.2}$$

Necessary conditions for the unbiasedness, sparsity, and continuity of the SCAD penalty have been proved by Antoniadis and Fan (2001). This penalty function involves two unknown parameters a and δ.

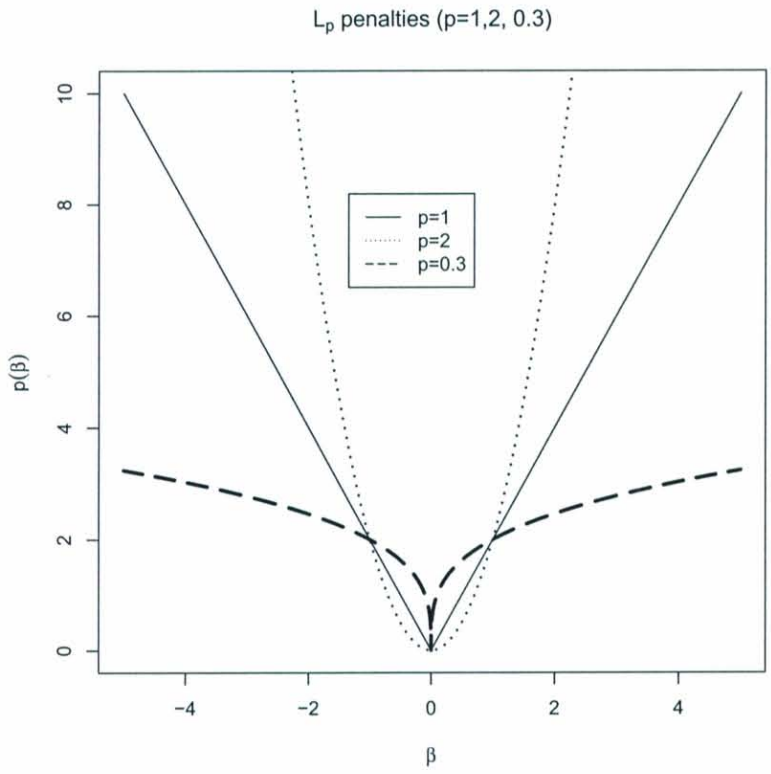

Figure 2.1: L_p penalty function

As shown in Figs. 2.1 and 2.2, all the penalty functions are singular at the origin, satisfying $p_\delta(0+) > 0$. This is the necessary condition for sparsity in variable selection. As shown in Fig. 2.2, the HARD and SCAD penalties are constant when β is large, indicating that there is no excessive penalization for large regression coefficients. However, SCAD is smoother than HARD and hence yields a continuous

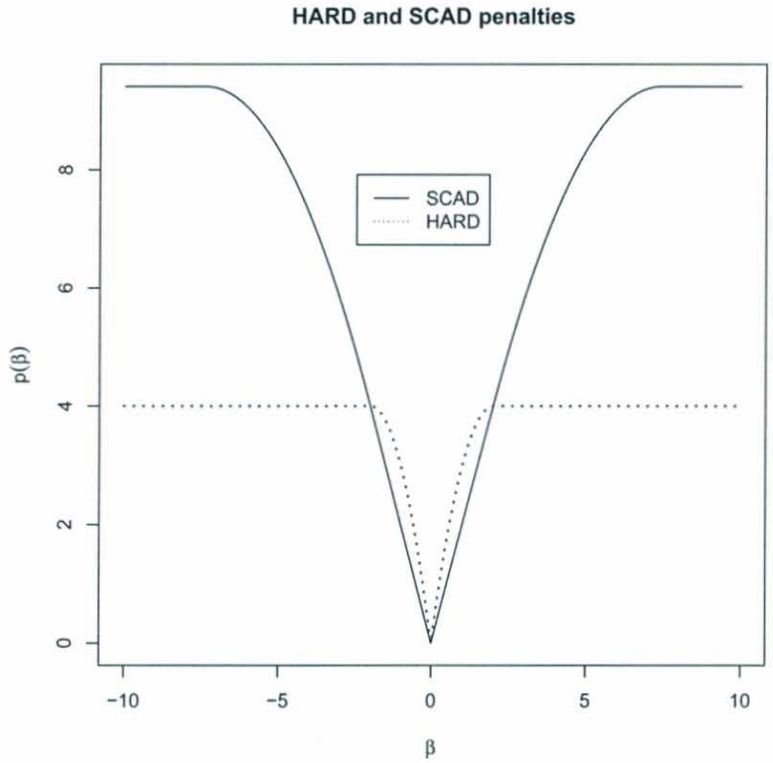

Figure 2.2: SCAD and HARD penalty functions

estimator.

Let $\boldsymbol{\beta}_0 = \left(\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T\right)^T$ be the true value of $\boldsymbol{\beta}$. Without loss of generality, we assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$ and all components of $\boldsymbol{\beta}_{10}$ are nonzero. Let $I(\boldsymbol{\beta}_0)$ be the Fisher information matrix and let $I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$ be the Fisher information given $\boldsymbol{\beta}_{20} = \mathbf{0}$. Under some regularity conditions, Fan and Li (2001) show that the estimate of the regression parameter based on the SCAD penalty, $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T\right)^T$, satisfies the oracle properties

for a certain choice of tuning parameter (δ, a), since

$$\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} \mathbf{0} \text{ and } \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{D} N\left(\mathbf{0},\ \mathrm{I}_1^{-1}(\boldsymbol{\beta}_{10}, \mathbf{0})\right).$$

The SCAD penalty function involves two unknown parameters, δ and a. In practice, we could search for the best pair (δ, a) over a two-dimensional structure using cross-validation (CV) or generalized cross-validation (GCV; Craven and Wahba, 1979). However, this would be computationally expensive. From a Bayesian point of view, Fan and Li (2001) suggested setting $a = 3.7$ and using GCV to select the best value of δ.

2.1 Local Quadratic Approximations and Standard Errors

The penalty function $p_\delta(|\beta_j|)$ is irregular at the origin and does not have continuous second-order derivatives at some points. Special care is needed in the application of the Newton-Raphson algorithm. Fan and Li (2001) locally approximate the SCAD penalty function by quadratic functions as follows. Suppose our initial value $\boldsymbol{\beta}_0$ is close to the maximizer of (2.1). If β_{j0} is very close to zero, then set $\hat{\beta}_j = 0$, otherwise, the penalty $p_\delta(|\beta_j|)$ can be locally approximated by the quadratic functions via

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

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

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

A disadvantage of this approximation is that once a coefficient has been shrunk to zero, it will stay at zero. However, this method significantly reduces the computational burden. Now we assume that the first two partial derivatives of the log-likelihood function are continuous, so that it is a smooth function with respect to $\boldsymbol{\beta}$. The first term in (2.1) can be locally approximated by a quadratic function via Taylor's expansion. The maximization problem (2.1) can be reduced to a quadratic maximization problem and the Newton-Raphson algorithm can be used. Therefore, (2.1) can be locally approximated by

$$\ell(\boldsymbol{\beta}_0) + \Delta\ell(\boldsymbol{\beta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\Delta^2\ell(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) - \frac{1}{2}n\boldsymbol{\beta}^T\Sigma_\delta(\boldsymbol{\beta}_0)\boldsymbol{\beta}, \quad (2.3)$$

where $\Delta\ell(\boldsymbol{\beta}_0) = \dfrac{\partial\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}}$, $\Delta^2\ell(\boldsymbol{\beta}_0) = \dfrac{\partial^2\ell(\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}$.

The quadratic maximization problem (2.3) is solved via the Newton-Raphson algorithm. In this algorithm, the update at the $(k+1)^{th}$ iteration is

$$\boldsymbol{\beta}^{k+1} = \boldsymbol{\beta}^k - \left[\Delta^2\ell(\boldsymbol{\beta}^k) - n\Sigma_\delta(\boldsymbol{\beta}^k)\right]^{-1}\left[\Delta\ell(\boldsymbol{\beta}^k) - n\mathrm{U}_\delta(\boldsymbol{\beta}^k)\right]$$

where $\Sigma_\delta(\boldsymbol{\beta}^k) = \text{diag}\left[\dfrac{p'_\delta(|\beta_1^k|)}{|\beta_1^k|}, \ldots, \dfrac{p'_\delta(|\beta_p^k|)}{|\beta_p^k|}\right]$ and $\mathrm{U}_\delta(\boldsymbol{\beta}^k) = \Sigma_\delta(\boldsymbol{\beta}^k)\boldsymbol{\beta}^k$.

The sandwich formula for the standard errors of the estimated parameters exists immediately because this method estimates the parameters and selects the variables at the same time. The standard errors of the estimated parameters are given by

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left[\Delta^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\delta(\hat{\boldsymbol{\beta}})\right]^{-1} \widehat{\text{cov}}\left\{\Delta \ell(\hat{\boldsymbol{\beta}})\right\} \left[\Delta^2 \ell(\hat{\boldsymbol{\beta}}) - n\Sigma_\delta(\hat{\boldsymbol{\beta}})\right]^{-1}.$$

Fan and Li (2001) conducted a series of Monte-Carlo simulations in linear regression, robust regression, and logistic regression and showed that the penalized-likelihood variable selection using the SCAD penalty performs better than the LASSO, HARD, and information-theoretic approaches.

Chapter 3

Variable Selection via Penalized Empirical Likelihood

The empirical likelihood method is a powerful inference tool with promising applications in many areas of statistics. In this chapter, we briefly introduce the basic concept of empirical likelihood. We then discuss the penalized-empirical-likelihood based variable selection method.

3.1 Empirical Likelihood (EL)

We first outline the empirical likelihood as discussed by Owen (1988, 1990). For a given random sample $y_1, y_2, \ldots, y_n$ from an unknown distribution function $F(y)$, the empirical likelihood function of F is defined to be

$$L_n(F) = \prod_{i=1}^{n} p_i,$$

where $p_i = F(\{y_i\}) = Pr(Y_i = y_i)$. The empirical likelihood is maximized without any further information about the empirical distribution function F

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \leq y),$$

where $I(*)$ is the indicator function and the inequality is expressed componentwise. In general, it is more common to work with the empirical log-likelihood

$$l_n(F) = \sum_{i=1}^{n} \log(p_i), \tag{3.1}$$

subject to the constraints $\sum_{i=1}^{n} p_i = 1$ and $p_i > 0, i = 1, 2, \ldots, n$. Suppose we want to investigate inference on the parameters under the assumption that F is a member of a nonparametric distribution family $\mathcal{F}$, say $\mu = T(F)$ for some functional T of the distribution. Inference for parameter μ can be obtained using the likelihood approach, if we know the likelihood value at μ. For a given value of μ, the population $F \in \mathcal{F}$ is such that $T(F) = \mu$. The task is to choose the F that best represents μ. The

notion of profile likelihood is to find the F at which the empirical likelihood attains the maximum value among the set of $T(F) = \mu$. The profile empirical likelihood function is defined to be

$$L_n(\mu) = \sup \{L_n(F) \mid T(F) = \mu,\ F \in \mathcal{F}\}.$$

We can construct the likelihood inference on μ based on $L_n(\mu)$. This likelihood has similar properties to its parametric counterpart. Since $L_n(\mu) \leq n^{-n}$, it is convenient to standardize $L_n(\mu)$ by defining the likelihood ratio function to be

$$R(F) = n^n L_n(\mu),$$

and it is easily shown that this can be written as

$$R(F) = \prod_{i=1}^{n} np_i.$$

The likelihood ratio function has a maximum value of 1. For simplicity, we can perform inference on any function F using the population mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)$, via the profile empirical likelihood. The profile empirical log-likelihood for $\boldsymbol{\mu}$ is defined to be

$$\ell(\boldsymbol{\mu}) = \sup \left\{ l_n(F) :\ p_i > 0, i = 1, 2, \ldots, n;\ \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i(y_i - \boldsymbol{\mu}) = 0 \right\}. \quad (3.2)$$

We can compute $\ell(\boldsymbol{\mu})$ by maximizing $\left\{\sum_{i=1}^{n}\log(p_i)\right\}$ by the Lagrange multiplier method under the above constraints. The Lagrange multiplier method is very effective for this constraint maximization problem. Define

$$G(p_1, p_2, \ldots p_n, \boldsymbol{\lambda}, \gamma) = \sum_{i=1}^{n}\log(p_i) + \gamma\left[\sum_{i=1}^{n}p_i - 1\right] - n\boldsymbol{\lambda}^T\left[\sum_{i=1}^{n}p_i(y_i - \boldsymbol{\mu})\right],$$

where $\boldsymbol{\lambda}$ (vector-valued) and γ are Lagrange multipliers. By setting the partial derivative of G with respect to p_i to zero, we get

$$\hat{p}_i = \frac{1}{n\left\{1 + \hat{\boldsymbol{\lambda}}^T(y_i - \boldsymbol{\mu})\right\}}, \text{ for } i = 1, 2, \ldots, n,$$

and the Lagrange multiplier $\hat{\boldsymbol{\lambda}} = \hat{\lambda}(\mu)$ is the solution of

$$\sum_{i=1}^{n}\frac{(y_i - \boldsymbol{\mu})}{1 + \boldsymbol{\lambda}^T(y_i - \boldsymbol{\mu})} = 0.$$

Therefore, we can write the profile empirical likelihood function as

$$\ell(\boldsymbol{\mu}) = -n\log(n) - \sum_{i=1}^{n}\log(1 + \hat{\boldsymbol{\lambda}}^T(\mu)(y_i - \boldsymbol{\mu})).$$

Now we define the profile empirical log-likelihood ratio function to be

$$W(\boldsymbol{\mu}) = \sum_{i=1}^{n}\log(n\hat{p}_i) = \sum_{i=1}^{n}\log\left[1 + \hat{\boldsymbol{\lambda}}^T(\mu)(y_i - \boldsymbol{\mu})\right].$$

Owen (1990) showed that, when μ_0 is the true population mean, $2W(\mu_0) \xrightarrow{D} \chi^2_d$ as $n \longrightarrow \infty$, similar to the parametric likelihood ratio function of Wilks (1938). This result is useful for hypothesis tests on parameter μ and for the construction of $100(1-\alpha)\%$ confidence regions, defined by

$$\{\mu : 2W(\mu) \leq \chi_d^2(1-\alpha)\},$$

where $\chi_d^2(1-\alpha)$ is the $(1-\alpha)^{th}$ quantile of the chi-square distribution with d degrees of freedom. This is different from the confidence intervals based on a normal approximation.

3.2 Penalized Empirical Likelihood based Variable Selection

Owen (1991) first considered EL for linear models. EL confidence regions for regression coefficients in linear models were studied by Chen (1994). We consider a linear model of the following form

$$y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i, \ i = 1, 2, \ldots, n,$$

where $\boldsymbol{X}_i \in \mathcal{R}^p$ is a vector of covariates and $\boldsymbol{\beta} \in \mathcal{R}^p$ a vector of parameters. We assume that the $y_i|\boldsymbol{X}_i$s are conditionally independent. We also assume that the error term ϵ_i is independent and identically distributed with mean zero and finite variance σ^2. Thus, $E(y_i|\boldsymbol{X}_i) = \boldsymbol{X}_i^T\boldsymbol{\beta}$ is the conditional mean function and $\text{Var}(y_i|\boldsymbol{X}_i) = \sigma^2$.

Following Owen (1991) and Qin and Lawless (1994), we can extend the empirical likelihood inferences for linear models based on a set of estimating functions $g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})$. Assume that the generalized linear model is defined by $E\left[g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})\right] = 0$. In general, g is a vector of $p \times 1$ estimating functions. The profile empirical log-likelihood function of $\boldsymbol{\beta}$ is defined by

$$\ell(\boldsymbol{\beta}) = \sup\left[\sum_{i=1}^{n} \log(p_i) : \; p_i > 0, i = 1, 2, \ldots, n; \; \sum_{i=1}^{n} p_i = 1, \; \sum_{i=1}^{n} p_i g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) = 0\right].$$

Using the Lagrange multiplier method discussed in Section 3.1, we can define

$$G(p_1, p_2, \ldots p_n, \boldsymbol{\lambda}, \gamma) = \sum_{i=1}^{n} \log(p_i) + \gamma\left[\sum_{i=1}^{n} p_i - 1\right] - n\boldsymbol{\lambda}^T\left[\sum_{i=1}^{n} p_i g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})\right],$$

where $\boldsymbol{\lambda}$ (vector valued) and γ are Lagrange multipliers. Setting the partial derivative of G with respect to p_i equal to zero gives

$$\hat{p}_i = \frac{1}{n\left\{1 + \hat{\boldsymbol{\lambda}}^T g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})\right\}}, \text{ for } i = 1, 2, \ldots, n, \tag{3.3}$$

where the Lagrange multiplier $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}(\beta)$ is the solution of

$$\sum_{i=1}^{n} \frac{g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})}{1 + \boldsymbol{\lambda}^T g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})} = 0. \tag{3.4}$$

This leads to the profile empirical log-likelihood function

$$\ell(\boldsymbol{\beta}) = -n\log(n) - \sum_{i=1}^{n} \log(1 + \hat{\boldsymbol{\lambda}}^T(\beta) g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}))$$

and the profile empirical log-likelihood ratio function is defined to be

$$W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log(n\hat{p}_i) = \sum_{i=1}^{n} \log(1 + \hat{\boldsymbol{\lambda}}^T(\beta) g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})). \tag{3.5}$$

Now we define the penalized empirical likelihood estimator of $\boldsymbol{\beta}$ as the maximizer of

$$\begin{aligned} \mathbf{L}(\boldsymbol{\beta}) &= -n\log(n) - \sum_{i=1}^{n} \left[\log(1 + \hat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta}) g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}))\right] - n\sum_{j=1}^{p} p_\delta(|\beta_j|) \\ &= \ell(\boldsymbol{\beta}) - n\sum_{j=1}^{p} p_\delta(|\beta_j|) \end{aligned} \tag{3.6}$$

with respect to $\boldsymbol{\beta}$, where $p_\delta(*)$ is the penalty function. We can use any of the penalty functions discussed in Chapter 2. Variyath (2006) first introduced the PEL, but reported some computational issues with over-penalizations. We use the continuous differential smoothly clipped absolute deviation (SCAD) penalty function with two unknown tuning parameters (δ, a) proposed by Fan and Li (2001) and defined in (2.2). In the next section we will discuss the distribution properties of the penalized empirical likelihood estimates of $\hat{\boldsymbol{\beta}}$ derived by Variyath (2006). The algorithm for the penalized empirical likelihood will be discussed in the next chapter.

3.3 Distributional Properties

Variyath (2006) stated and proved theorems in connection with PEL; we reproduce them here. Let $\boldsymbol{\beta}_0 = \left(\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T\right)^T$ be the true value of $\boldsymbol{\beta}$ with vector lengths of k

and $p-k$ respectively. Without loss of generality, we assume that $\boldsymbol{\beta}_{20} = \mathbf{0}$ and all components of $\boldsymbol{\beta}_{10}$ are nonzero. Let $I(\boldsymbol{\beta}_0)$ be the Fisher information matrix and let $I_1(\boldsymbol{\beta}_{10}, \mathbf{0})$ be the Fisher information given $\boldsymbol{\beta}_{20} = \mathbf{0}$. Under some regularity conditions, our penalized empirical likelihood SCAD estimator $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T\right)^T$ satisfies the oracle properties for a certain choice of the tuning parameters (δ, a). Hence, it is easy to prove that

$$\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} \mathbf{0} \quad \text{and} \quad \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{D} N\left(\mathbf{0},\ \mathrm{I}_1^{-1}(\boldsymbol{\beta}_{10}, \mathbf{0})\right).$$

The following theorem proves the existence of a local maximizer of the penalized empirical likelihood $\mathbf{L}(\boldsymbol{\beta})$.

Theorem 3.3.1 *(Variyath, 2006) Suppose $(y_i, \boldsymbol{X}_i), i = 1, 2, \ldots, n$ is a set of independent and identically distributed random vectors. Let $g_i(\boldsymbol{\beta}) = g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})$ be the estimating functions for $\boldsymbol{\beta} \in \mathcal{R}^p$ such that for each $i = 1, 2, \ldots n$,*

$$E\left\{g_i(\boldsymbol{\beta_0})\right\} = 0$$

for some $\boldsymbol{\beta}_0$. Also assume that

(i) $V = E\left\{g(\boldsymbol{\beta_0})g^T(\boldsymbol{\beta_0})\right\}$ is positive definite,

(ii) $\dfrac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$ is continuous in $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta_0}$,

(iii) the rank of $E\left\{\dfrac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right\}$ is p in a neighborhood of $\boldsymbol{\beta_0}$,

(iv) there exists some functions $G(\boldsymbol{y}, \boldsymbol{X})$ *such that in a neighborhood of* $\boldsymbol{\beta_0}$,

$$\left|\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right| < G(\boldsymbol{y}, \boldsymbol{X}),\ \|g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})\|^3 < G(\boldsymbol{y}, \boldsymbol{X})$$

such that $E\left[G(\boldsymbol{y}, \boldsymbol{X})\right] < \infty$. *The tuning parameter* δ *is chosen as a function of* n *such that* $max(p''_{\delta_n}|\beta_{jo}| : \beta_{jo} \neq 0) \to 0$ *as* $n \to \infty$. *Then there exists a local maximizer* $\hat{\boldsymbol{\beta}}$ *of* $\mathbf{L}(\boldsymbol{\beta})$ *such that* $\left\|\hat{\boldsymbol{\beta}} - \beta_0\right\| = O_p(n^{-1/2} + b_n)$, *where* $b_n = \max(p'_{\delta_n}|\beta_{jo}| : \beta_{jo} \neq 0)$.

Theorem 3.3.1 shows that for an appropriate choice of δ_n, there exists a root-n consistent penalized empirical likelihood estimator. The following lemma shows that this estimator must have the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$.

Lemma 3.3.2 *(Variyath, 2006) Suppose* $(y_i, \boldsymbol{X}_i), i = 1, 2, \ldots, n$ *is a set of independent and identically distributed random vectors. Let* $g_i(\boldsymbol{\beta}) = g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})$ *be the estimating function for* $\boldsymbol{\beta} \in \mathcal{R}^p$ *such that, for each* $i = 1, 2, \ldots, n$,

$$E\left\{g_i(\boldsymbol{\beta_0})\right\} = 0$$

for some $\boldsymbol{\beta}_0$. *Also assume that*

(i) $V = E\left\{g(\boldsymbol{\beta_0})g^T(\boldsymbol{\beta_0})\right\}$ *is positive definite,*

(ii) $\dfrac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$ *is continuous in* $\boldsymbol{\beta}$ *in a neighborhood of* $\boldsymbol{\beta_0}$,

(iii) the rank of $E\left\{\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right\}$ *is* p *in a neighborhood of* $\boldsymbol{\beta_0}$,

(iv) there exists some functions $G(\boldsymbol{y}, \boldsymbol{X})$ *such that in a neighborhood of* $\boldsymbol{\beta_0}$,

$$\left|\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right| < G(\boldsymbol{y}, \boldsymbol{X}),\ \|g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})\|^3 < G(\boldsymbol{y}, \boldsymbol{X})$$

such that $E\left[G(\boldsymbol{y}, \boldsymbol{X})\right] < \infty$.

Assume that

$$\lim_{n\to\infty} \lim_{\beta\to 0+} \left\{\frac{p'_{\delta_n}(\boldsymbol{\beta})}{\delta_n}\right\} > 0. \tag{3.7}$$

If $\delta_n \to 0$ *and* $\sqrt{n}\delta_n \to \infty$, *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,

$$\mathbf{L}\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ 0 \end{pmatrix}\right\} = max_{\|\beta_2\| \le Cn^{-1/2}} \mathbf{L}\left\{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\right\}.$$

Using the above lemma, one can prove the following theorem on the asymptotic normality of the empirical likelihood estimate.

Theorem 3.3.3 *(Variyath, 2006) In addition to the conditions of Theorem 3.3.1 and Lemma 3.3.2, suppose that* $\frac{\partial^2 g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}}$ *is continuous in* $\boldsymbol{\beta}$ *in a neighborhood of the true value of* $\boldsymbol{\beta}_0$ *and is bounded by some integrable function* $G(\boldsymbol{y}, \boldsymbol{X})$. *Then*

$$\sqrt{n}\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (-\boldsymbol{\Delta})^{-1} p'_\delta(|\boldsymbol{\beta}|)\right\} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Delta})$$

where $\hat{\boldsymbol{\beta}}$ *is the penalized empirical likelihood estimate of* $\boldsymbol{\beta}$ *and*

$$\boldsymbol{\Delta} = \left[E\left\{ \frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \right\}^T \left\{ E\left\{ g(\boldsymbol{\beta}_0) g^T(\boldsymbol{\beta}_0) \right\}^{-1} \right\} E\left\{ \frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \right\} \right]^{-1}.$$

3.4 Penalized Adjusted Empirical Likelihood

Computation of $W(\boldsymbol{\beta})$ for a given value of $\boldsymbol{\beta}$ may lead to some technical problem. The solution for $\boldsymbol{\lambda}$ must satisfy $\left\{ 1 + \hat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta}) g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}) \right\} > 0$ for all $i = 1, \ldots, n$. A necessary and sufficient condition for its existence is that the vector '$\mathbf{0}$' is an inner point of the convex hull of $\{g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}), i = 1, \ldots, n\}$. The true parameter value $\boldsymbol{\beta}_0$ is the unique solution of $E[g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})] = 0$. But, under some moment conditions on $g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})$ (Owen, 2001), the convex hull $\{g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}), i = 1, \ldots, n\}$ contains 0 as its inner point with probability 1 as $n \to \infty$. When $\boldsymbol{\beta}$ is not close to $\boldsymbol{\beta}_0$, or when n is small, there is a considerable chance that the solution of (3.4) does not exists. To avoid this problem, Chen, Variyath and Abraham (2008) introduced the adjusted empirical likelihood.

Denote $g_i(\boldsymbol{\beta}) = g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})$ and $\overline{g}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} g_i(\boldsymbol{\beta})$ for any given $\boldsymbol{\beta}$. For some positive constant a_n, define

$$\begin{aligned} g_{n+1}(\boldsymbol{\beta}) &= -\frac{a_n}{n} \sum_{i=1}^{n} g_i(\boldsymbol{\beta}) \\ &= -a_n \overline{g}_n(\boldsymbol{\beta}). \end{aligned}$$

Now the adjusted profile empirical log-likelihood ratio function is defined as

$$W^*(\boldsymbol{\beta}) = \sup\left[\sum_{i=1}^{n+1}\log\left[(n+1)p_i\right] : \; p_i > 0, i = 1, 2, \ldots, n+1; \; \sum_{i=1}^{n+1} p_i = 1, \; \sum_{i=1}^{n+1} p_i g_i(\boldsymbol{\beta}) = 0\right],$$
$$= \sum_{i=1}^{n+1}\log\left[1 + \hat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta})g_i(\boldsymbol{\beta})\right]$$

with $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}(\beta)$ being the solution of $\sum_{i=1}^{n+1}\frac{g_i(\boldsymbol{\beta})}{1+\boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})} = 0$. Note that now 0 always lies inside the convex hull of $\{g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta}), i = 1, \ldots, n\}$. The adjusted empirical log-likelihood ratio function is well defined after adding a pseudo-value $g_{n+1}(\boldsymbol{\beta})$. For a wide range of a_n, $W^*(\boldsymbol{\beta})$ have same first order asymptotic properties of $W(\boldsymbol{\beta})$ (see Chen et al., 2008). We extend this idea of penalized adjusted empirical likelihood to avoid the technical problem of non-existence of solution to (3.4) for any given value of $\boldsymbol{\beta}$.

Now we define the penalized adjusted empirical likelihood estimator of $\boldsymbol{\beta}$ as the maximizer of

$$\mathbf{L}^*(\boldsymbol{\beta}) = -(n+1)\log(n+1) - \sum_{i=1}^{n+1}\left[\log(1 + \hat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta})g_i(\boldsymbol{\beta}))\right] - (n+1)\sum_{j=1}^{p} p_\delta(|\beta_j|)$$
$$= \ell^*(\boldsymbol{\beta}) - (n+1)\sum_{j=1}^{p} p_\delta(|\beta_j|) \tag{3.8}$$

with respect to $\boldsymbol{\beta}$, where $p_\delta(*)$ is the penalty function defined in (2.2). This adjustment is particularly useful because even for some undesirable values of $\boldsymbol{\beta}$ and tuning parameters, the proposed algorithm guarantees a solution. Now, we can show that the penalized adjusted empirical likelihood has the same asymptotic properties as

the penalized empirical likelihood detailed in Section 3.3. We state and prove the following theorems and lemma to show that the penalized adjusted empirical likelihood estimates have oracle properties.

Theorem 3.4.1 *Suppose* $(y_i, \boldsymbol{X_i}), i = 1, 2, \ldots, n$ *is a set of independent and identically distributed random vectors. Let* $g_i(\boldsymbol{\beta}) = g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})$ *be the estimating functions for* $\boldsymbol{\beta} \in \mathcal{R}^p$ *such that for each* $i = 1, 2, \ldots n$,

$$E\left\{g_i(\boldsymbol{\beta_0})\right\} = 0$$

for some $\boldsymbol{\beta}_0$. *Let* $\overline{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta})$ *and* $g_{n+1}(\boldsymbol{\beta}) = -a_n\overline{g}_n(\boldsymbol{\beta})$, *where* a_n *is a positive constant. Also assume that*

(i) $V = E\left\{g(\boldsymbol{\beta_0})g^T(\boldsymbol{\beta_0})\right\}$ *is positive definite,*

(ii) $\dfrac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$ *is continuous in* $\boldsymbol{\beta}$ *in a neighborhood of* $\boldsymbol{\beta_0}$,

(iii) the rank of $E\left\{\dfrac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right\}$ *is* p *in a neighborhood of* $\boldsymbol{\beta_0}$,

(iv) there exists some functions $G(\boldsymbol{y}, \boldsymbol{X})$ *such that in a neighborhood of* $\boldsymbol{\beta_0}$,

$$\left|\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right| < G(\boldsymbol{y}, \boldsymbol{X}),\ \|g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})\|^3 < G(\boldsymbol{y}, \boldsymbol{X})$$

such that $E\left[G(\boldsymbol{y}, \boldsymbol{X})\right] < \infty$. *The tuning parameter* δ *is chosen as a function of* m *such that* $\max(p''_{\delta_m}|\beta_{jo}| : \beta_{jo} \neq 0) \to 0$ *as* $m \to \infty$, *where* $m = n + 1$. *Then*

there exists a local maximizer $\hat{\boldsymbol{\beta}}$ *of* $\mathbf{L}^*(\boldsymbol{\beta})$ *such that* $\left\|\hat{\boldsymbol{\beta}} - \beta_0\right\| = O_p(m^{-1/2} + b_m)$, *where* $b_m = \max(p'_{\delta_n}|\beta_{jo}| : \beta_{jo} \neq 0)$.

Proof

Let $\alpha_m = m^{-1/2} + b_m$. It is sufficient to show that for any $\epsilon > 0$, there exists a large enough C such that

$$\Pr\{\sup \mathbf{L}^*[(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}); \|\boldsymbol{u}\| = C] < \mathbf{L}^*(\boldsymbol{\beta}_0)\} \geq 1 - \epsilon. \tag{3.9}$$

This implies that for large m with probability at least $1 - \epsilon$, there exists a local maximizer in the ball $[(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}); \|\boldsymbol{u}\| = C]$. Hence, there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\alpha_m)$. Let

$$D_m^*(\boldsymbol{u}) = \mathbf{L}^*(\boldsymbol{\beta_0} + \alpha_m \boldsymbol{u}) - \mathbf{L}^*(\boldsymbol{\beta}_0).$$

Then

$$\begin{aligned} D_m^*(\boldsymbol{u}) &= \{\ell^*(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}) - \ell^*(\boldsymbol{\beta}_0)\} - \{p_\delta(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}) - p_\delta(\boldsymbol{\beta}_0)\} \\ &= \{\ell^*(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}) - \ell^*(\boldsymbol{\beta}_0)\} - m \sum_{j=1}^{k} \{p_\delta(|\boldsymbol{\beta}_{j0} + \alpha_m \boldsymbol{u}|) - p_\delta(|\boldsymbol{\beta}_{j0}|)\}, \end{aligned}$$

where k is the number of components in $\boldsymbol{\beta}_{10}$. The Lagrange multiplier in $\boldsymbol{\lambda}(\beta_0)$ can be expressed as

$$\begin{aligned} \boldsymbol{\lambda}(\beta_0) &= \left\{ \frac{1}{m} \sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0) g_i^T(\boldsymbol{\beta}_0) \right\}^{-1} \left\{ \frac{1}{m} \sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0) \right\} + o(m^{-1/2}) \\ &= V_m^{-1}(\boldsymbol{\beta_0}) \overline{g}_m(\boldsymbol{\beta}_0) + o(m^{-1/2}) = O_p(m^{-1/2}), \end{aligned}$$

where $\overline{g}_m(\boldsymbol{\beta}_0) = \frac{1}{m}\sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0)$ and $V_m(\boldsymbol{\beta}_0) = \frac{1}{m}\sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0)g_i^T(\boldsymbol{\beta}_0)$. Hence,

$$\begin{aligned}
-\ell^*(\boldsymbol{\beta}_0) &= \sum_{i=1}^{m} \log\left\{1 + \boldsymbol{\lambda}^T(\beta_0)g_i(\boldsymbol{\beta}_0)\right\} + o_p(1) \\
&= \sum_{i=1}^{m} \boldsymbol{\lambda}^T(\beta_0)g_i(\boldsymbol{\beta}_0) - \frac{1}{2}\sum_{i=1}^{m}\left[\boldsymbol{\lambda}^T(\beta_0)g_i(\boldsymbol{\beta}_0)\right]^2 + o_p(1) \\
&= \frac{m}{2}\overline{g}_m^T(\boldsymbol{\beta}_0)V_m^{-1}(\boldsymbol{\beta}_0)\overline{g}_m(\boldsymbol{\beta}_0) + o_p(1).
\end{aligned}$$

Similarly,

$$\begin{aligned}
-\ell^*(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u}) &= \sum_{i=1}^{m} \boldsymbol{\lambda}^T(\beta_0 + \alpha_m\boldsymbol{u})g_i(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u}) \\
&\quad -\frac{1}{2}\sum_{i=1}^{m}\left[\boldsymbol{\lambda}^T(\beta_0 + \alpha_m\boldsymbol{u})g_i(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u})\right]^2 + o_p(1) \\
&= \frac{m}{2}\left\{\frac{1}{m}\sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u})\right\}^T \left\{\frac{1}{m}\sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u})g_i^T(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u})\right\}^{-1} \\
&\quad \left\{\frac{1}{m}\sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u})\right\} + o_p(1) \\
&= \frac{m}{2}\left\{\overline{g}_m(\boldsymbol{\beta}_0) + \alpha_m\boldsymbol{u}\frac{1}{m}\sum_{i=1}^{m}\frac{\partial g_i(\boldsymbol{\beta}_0)}{\partial\beta}\right\}^T \left\{\frac{1}{m}\sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0)g_i^T(\boldsymbol{\beta}_0)\right\}^{-1} \\
&\quad \left\{\overline{g}_m(\boldsymbol{\beta}_0) + \alpha_m\boldsymbol{u}\frac{1}{m}\sum_{i=1}^{m}\frac{\partial g_i(\boldsymbol{\beta}_0)}{\partial\beta}\right\} + o_p(1) \\
&= \frac{m}{2}\left\{\overline{g}_m(\boldsymbol{\beta}_0) + \alpha_m\boldsymbol{u}E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]\right\}^T V_m^{-1}(\boldsymbol{\beta}_0)\left\{\overline{g}_m(\boldsymbol{\beta}_0) + \alpha_m\boldsymbol{u}E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]\right\} + o_p(1).
\end{aligned}$$

Now

$$\ell^*(\boldsymbol{\beta}_0 + \alpha_m\boldsymbol{u}) - \ell^*(\boldsymbol{\beta}_0) = -m\alpha_m\boldsymbol{u}^T\left\{E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]^T E\left[g(\boldsymbol{\beta}_0)g^T(\boldsymbol{\beta}_0)\right]^{-1}\overline{g}_m(\boldsymbol{\beta}_0)\right\}$$

$$-\frac{1}{2}m\alpha_m^2\boldsymbol{u}^T\left\{E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]^T V_m^{-1}(\boldsymbol{\beta}_0)E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]\right\}\boldsymbol{u}+o_p(1).$$

Now, letting

$$\Delta=\left\{E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]^T E\left[g(\boldsymbol{\beta}_0)g^T(\boldsymbol{\beta}_0)\right]^{-1}E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]\right\},$$

we have

$$D_m^*(\boldsymbol{u})\le -m\alpha_m\boldsymbol{u}^T\left\{E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial\beta}\right]^T\Delta\overline{g}_m(\boldsymbol{\beta}_0)\right\}-\frac{m}{2}\alpha_m^2\boldsymbol{u}^T\Delta^{-1}\boldsymbol{u} \tag{3.10}$$

$$-\sum_{i=1}^{k}\left\{m\alpha_m p'_{\delta_m}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})u_j+m\alpha_m^2 p''_{\delta_m}(|\beta_{j0}|)u_j^2[1+o(1)]\right\}.$$

It can easily be shown that Δ is the asymptotic variance of $\sqrt{m}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)$, and so the representation is similar to normalized parametric likelihood. By the central limit theorem, $\overline{g}_m(\boldsymbol{\beta}_0)$ is $O_p(m^{-1/2})$, thus the first term on the right-hand side of (3.10) is of order $O_p(m^{1/2}\alpha_m)=O_p(m\alpha_m^2)$. By selecting a large C, the second term dominates the first term uniformly in $\|\boldsymbol{u}\|=C$. The third term is bounded by

$$\sqrt{k}m\alpha_m b_m\|\boldsymbol{u}\|+m\alpha_m^2\max\left\{|p''_\delta(|\beta_{j0}|)|:\beta_{j0}\neq 0\right\}\|\boldsymbol{u}\|^2.$$

This is also dominated by the second term in (3.10). Hence, by choosing a sufficiently large value of C, (3.9) holds. This completes the proof. Theorem 3.4.1 shows that for an appropriate choice of δ_m, there exists a root-m consistent penalized empirical

likelihood estimator. The following lemma shows that this estimator must have the sparsity property $\hat{\boldsymbol{\beta}}_2 = 0$.

Lemma 3.4.2 *Suppose* $(y_i, \boldsymbol{X}_i), i = 1, 2, \ldots, n$ *is a set of independent and identically distributed random vectors. Let* $g_i(\boldsymbol{\beta}) = g(y_i, \boldsymbol{X}_i, \boldsymbol{\beta})$ *be the estimating function for* $\boldsymbol{\beta} \in \mathcal{R}^p$ *such that, for each* $i = 1, 2, \ldots, n$,

$$E\left\{g_i(\boldsymbol{\beta_0})\right\} = 0$$

for some $\boldsymbol{\beta}_0$. *Let* $\overline{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta})$ *and* $g_{n+1}(\boldsymbol{\beta}) = -a_n\overline{g}_n(\boldsymbol{\beta})$, *where* a_n *is a positive constant. Also assume that*

(i) $V = E\left\{g(\boldsymbol{\beta_0})g^T(\boldsymbol{\beta_0})\right\}$ *is positive definite,*

(ii) $\dfrac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}$ *is continuous in* $\boldsymbol{\beta}$ *in a neighborhood of* $\boldsymbol{\beta_0}$,

(iii) the rank of $E\left\{\dfrac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right\}$ *is* p *in a neighborhood of* $\boldsymbol{\beta_0}$,

(iv) there exists some functions $G(\boldsymbol{y}, \boldsymbol{X})$ *such that in a neighborhood of* $\boldsymbol{\beta_0}$,

$$\left|\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right| < G(\boldsymbol{y}, \boldsymbol{X}),\ \|g(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta})\|^3 < G(\boldsymbol{y}, \boldsymbol{X})$$

such that $E\left[G(\boldsymbol{y}, \boldsymbol{X})\right] < \infty$.

Assume that

$$\lim_{m\to\infty}\lim_{\beta\to 0+}\left\{\frac{p'_{\delta_m}(\boldsymbol{\beta})}{\delta_m}\right\} > 0, \tag{3.11}$$

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

$$\mathbf{L}^*\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\0\end{pmatrix}\right\} = max_{\|\boldsymbol{\beta}_2\|\leq Cm^{-1/2}}\mathbf{L}^*\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\\boldsymbol{\beta}_2\end{pmatrix}\right\}.$$

Proof

Following Fan and Li (2001) in proving this Lemma, it is sufficient to show that for $\boldsymbol{\beta}$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(m^{-1/2})$ and for some small $\epsilon_n = Cm^{-1/2}$, and $j = k+1, \ldots, p$,

$$\frac{\partial \mathbf{L}^*(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} < 0 \qquad \text{for } 0 < \beta_j < \epsilon_m$$

$$\text{for } -\epsilon_m < \beta_j < 0. \tag{3.12}$$

Due to the condition on $p_{\delta_m}(|\boldsymbol{\beta}|)$, the task is equivalent to showing that, uniformly in $\boldsymbol{\beta}$,

$$\left|\frac{\partial \ell^*(\boldsymbol{\beta})}{\partial \beta_j}\right| = O_p(m^{1/2}).$$

That is, the slope around the true value of $\boldsymbol{\beta}$ is low compared to the slope of the penalty. Now

$$\ell^*(\beta_j) = -\sum_{i=1}^{m} \log\left\{1 + \lambda^T(\beta_j) g_i(\beta_j)\right\}$$

where we regard $\boldsymbol{\lambda}$ and g_i as functions of a specific component of $\boldsymbol{\beta}$ for simplicity. Note that

$$\frac{\partial \ell^*(\beta_j)}{\partial \beta_j} = -\sum_{i=1}^{m} \frac{1}{1+\lambda^T(\beta_j)g_i(\beta_j)} \left\{ \frac{\partial g_i(\beta_j)}{\partial \beta_j} \right\} \lambda(\beta_j).$$

Since $\beta_1 - \beta_{10} = O_p(m^{-1/2})$, it is simple to show that we still have

$$\max_{1\le i\le m} \|g_i(\beta_j)\| = o_p(m^{1/3}) \text{ and } \|\lambda^T(\beta_j)\| = O_p(m^{-1/2}).$$

Hence,

$$\lambda^T(\beta_j)g_i(\beta_j) = o_p(1)$$

uniformly in both $i = 1, 2, \ldots, m$ and $\boldsymbol{\beta}$. Thus we have

$$\begin{aligned} \left| \frac{\partial \ell(\beta)}{\partial \beta_j} \right| &\le \|\lambda^T(\beta_j)\| \sum_{i=1}^{m} \left\| \frac{\partial g_i(\beta_j)}{\partial \beta_j} \right\| [1 + o_p(1)] \\ &= O_p(m^{-1/2}) O_p(m)[1 + o_p(1)] \\ &= O_p(m^{1/2}). \end{aligned}$$

Using the above results, for each component of $\boldsymbol{\beta}$ we have

$$\frac{\partial \mathbf{L}^*(\boldsymbol{\beta})}{\partial \beta_j} = m\delta_m \left\{ -\delta_m^{-1} p'_{\delta_m}(|\beta_j|)\text{sgn}(\beta_j) + \delta_m^{-1} O_p(m^{-1/2}) \right\}.$$

Using the assumption (3.11), $\sqrt{m}\delta_m \to \infty$ and $\delta_m \to 0$, the sign of the derivative is completely determined by that of β_j. Hence (3.12) holds. This completes the proof. Using the above lemma, we can prove the following theorem on the asymptotic normality of the adjusted empirical likelihood estimate.

Theorem 3.4.3 *In addition to the conditions of Theorem 3.4.1 and Lemma 3.4.2, suppose the second derivatives of each component of g, say $g[k]$, $\frac{\partial^2 g[k]}{\partial \beta^2}$, a $p \times p$ matrix with the $(ij)^{th}$ entry $\frac{\partial^2 g[k]}{\partial \beta_i \partial \beta_j}$, is continuous in $\boldsymbol{\beta}$ in a neighborhood of $\boldsymbol{\beta}_0$ and is bounded by some integrable function $G(\boldsymbol{y}, \boldsymbol{X})$. Then*

$$\sqrt{m}\left\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (-\boldsymbol{\Delta})^{-1} p'_\delta(|\boldsymbol{\beta}|)\right\} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Delta})$$

where $\hat{\boldsymbol{\beta}}$ is the penalized empirical likelihood estimate of $\boldsymbol{\beta}$ and

$$\boldsymbol{\Delta} = \left[E\left\{\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}\right\}^T \left\{E\left\{g(\boldsymbol{\beta}_0) g^T(\boldsymbol{\beta}_0)\right\}^{-1}\right\} E\left\{\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}\right\}\right]^{-1}.$$

Proof

Due to the sparsity property given in Lemma 3.4.2, it is seen that the penalized adjusted empirical likelihood estimator with proper tuning parameter δ_n maximizes $\mathbf{L}^*\{(\boldsymbol{\beta}_1, \mathbf{0})^T\}$ with respect to $\boldsymbol{\beta}_1$. Hence,

$$\frac{\partial \mathbf{L}^*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})}{\partial \boldsymbol{\lambda}} = \mathbf{L}^*_{1,m}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}) = 0, \frac{\partial \mathbf{L}^*(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})}{\partial \boldsymbol{\beta}} = \mathbf{L}^*_{2,m}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}) = 0$$

where

$$\mathbf{L}^*_{1,m}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{m} \sum_{i=1}^{m} \frac{g_i(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}^T(\beta) g_i(\boldsymbol{\beta})}$$

and

$$\mathbf{L}^*_{2,m}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 + \boldsymbol{\lambda}^T(\beta) g_i(\boldsymbol{\beta})} \left(\frac{\partial g_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\lambda} + m p'_\delta(|\boldsymbol{\beta}_1|) \mathrm{sgn}(\boldsymbol{\beta}_1).$$

For notational simplicity, we do not differentiate $\frac{\partial \mathbf{L}^*}{\partial \boldsymbol{\beta}_1}$ and $\frac{\partial \mathbf{L}^*}{\partial \boldsymbol{\beta}}$ for the rest of the proof. That is, we present our proof as if $k = p$. If we expand these functions at $(\boldsymbol{\beta} = \boldsymbol{\beta}_0, \boldsymbol{\lambda} = \mathbf{0})$, we have

$$\begin{aligned}
\mathbf{L}^*_{1,m}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}) &= \mathbf{L}^*_{1,m}(\boldsymbol{\beta}_0, \mathbf{0}) \\
&\quad + \left[\frac{\mathbf{L}^*_{1,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\beta}}\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \left[\frac{\mathbf{L}^*_{1,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}^T}\right](\hat{\boldsymbol{\lambda}} - \mathbf{0}) + o_p(\delta_m) = 0, \\
\mathbf{L}^*_{2,m}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}) &= \mathbf{L}^*_{2,m}(\boldsymbol{\beta}_0, \mathbf{0}) \\
&\quad + \left[\frac{\mathbf{L}^*_{2,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\beta}}\right](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \left[\frac{\mathbf{L}^*_{2,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}^T}\right](\hat{\boldsymbol{\lambda}} - \mathbf{0}) + o_p(\delta_m) = 0
\end{aligned}$$

where $\delta_m = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\hat{\boldsymbol{\lambda}}\|$. The partial derivatives in the above expansions are

$$\begin{aligned}
\frac{\mathbf{L}^*_{1,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\beta}} &= \frac{1}{m}\sum_{i=1}^{n+1} \frac{\partial g_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \to -E\left\{\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\}, \\
\frac{\mathbf{L}^*_{1,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}} &= \frac{1}{m}\sum_{i=1}^{m} g_i(\boldsymbol{\beta}_0) g_i^T(\boldsymbol{\beta}_0) \to E\left\{g(\boldsymbol{\beta}_0) g^T(\boldsymbol{\beta}_0)\right\}, \\
\frac{\mathbf{L}^*_{2,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\beta}} &= p''_{\delta_m}(|\boldsymbol{\beta}_0|), \\
\frac{\mathbf{L}^*_{2,m}(\boldsymbol{\beta}_0, \mathbf{0})}{\partial \boldsymbol{\lambda}} &= \frac{1}{m}\sum_{i=1}^{m} \left\{\frac{\partial g_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\}^T \to E\left\{\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\}^T,
\end{aligned}$$

Therefore,

$$\begin{bmatrix} \hat{\boldsymbol{\lambda}} \\ \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} \end{bmatrix} = S_m^{-1} \begin{bmatrix} -\mathbf{L}^*_{1,m}(\boldsymbol{\beta}_1, \mathbf{0}) + o_p(\delta_m) \\ -\mathbf{L}^*_{2,m}(\boldsymbol{\beta}_1, \mathbf{0}) + o_p(\delta_m) \end{bmatrix}$$

with

$$S_m = \begin{bmatrix} -E\left\{g(\boldsymbol{\beta}_0) g^T(\boldsymbol{\beta}_0)\right\} & E\left\{\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\} \\ E\left\{\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right\}^T & p''_{\delta_m}(|\boldsymbol{\beta}_0|) \end{bmatrix}.$$

Since $\mathbf{L}^{*}_{1,m}(\boldsymbol{\beta}_{10}, \mathbf{0}) = \overline{g}_m(\boldsymbol{\beta}_0) = O_p(m^{-1/2})$, we can easily show that $\delta_m = O_p(m^{-1/2})$. When $p''_{\delta_m}(|\boldsymbol{\beta}|) \to \mathbf{0}$ as $m \to \infty$, the limiting distribution of $\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}$ will be asymptotically normal, i.e.,

$$\sqrt{m}\left\{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 + S^{22}_m \mathbf{L}_{2,m}(\boldsymbol{\beta}_1, \mathbf{0})\right\} \xrightarrow{D} N(\mathbf{0}, \Delta),$$

where

$$\Delta = \left\{ E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \beta}\right]^T \left\{E\left[g(\boldsymbol{\beta}_0) g^T(\boldsymbol{\beta}_0)\right]\right\}^{-1} E\left[\frac{\partial g(\boldsymbol{\beta}_0)}{\partial \beta}\right]\right\}^{-1},$$

and $S^{22}_m = -\Delta^{-1}$ is the $(2,2)^{th}$ element of S^{-1}_m assuming $p''_{\delta_m}(|\boldsymbol{\beta}|) = 0$. This completes the proof.

Chapter 4

Numerical Algorithm

To implement our method, we need an efficient numerical algorithm. Variyath (2006) reported some computational issues with over-penalizations that resulted in high bias. We maximize the PEL with respect to $\boldsymbol{\beta}$ using a modified Newton-Raphson algorithm. At each iteration of the Newton-Raphson method, we compute the Lagrange multiplier for an updated value of $\boldsymbol{\beta}$. Chen, Sitter, and Wu (2002) proposed a modified Newton-Raphson algorithm for computing the Lagrange multiplier for a given value of the parameter. This method is numerically stable, which is useful in this application. The numerical algorithm given in Section 4.1 and 4.2 can be easily extended to penalized adjusted empirical likelihood, by adding a pseudo-value $g_{n+1}(\boldsymbol{\beta}) = -a_n \overline{g}_n(\boldsymbol{\beta})$, where a_n is a positive constant.

4.1 Computation of Lagrange Multiplier

The Lagrange multiplier $\boldsymbol{\lambda}$ is estimated by solving the equation

$$\sum_{i=1}^{n} \frac{g_i(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})} = 0$$

for a given set of vectors $g_i(\boldsymbol{\beta})$, $i = 1, 2, \ldots n$. Note that the above equation is the derivative of R with respect to $\boldsymbol{\lambda}$ for a given $\boldsymbol{\beta}$, where

$$R = \sum_{i=1}^{n} \log \left\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\right\}. \tag{4.1}$$

In the empirical likelihood problem, the solution must satisfy the condition that

$$1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta}) > 0, \quad i = 1, 2, \ldots n.$$

The modified Newton-Raphson algorithm for estimating $\boldsymbol{\lambda}$ for a given value of $\boldsymbol{\beta}$ is as follows:

1. Set $\boldsymbol{\lambda}^c = 0$, $c = 0$, $\gamma^k = 1$, $\epsilon = 1e - 08$, and $\boldsymbol{\beta} = \boldsymbol{\beta}^0$.

2. Let $R^{\boldsymbol{\lambda}}$ and $R^{\boldsymbol{\lambda}\boldsymbol{\lambda}}$ be the first and second partial derivatives of R given in (4.1) with respect to $\boldsymbol{\lambda}$, which are given by

$$R^{\boldsymbol{\lambda}} = \sum_{i=1}^{n} \left[\frac{g_i(\boldsymbol{\beta})}{\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\}}\right], \quad R^{\boldsymbol{\lambda}\boldsymbol{\lambda}} = -\sum_{i=1}^{n} \left[\frac{g_i(\boldsymbol{\beta}) g_i^T(\boldsymbol{\beta})}{\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\}^2}\right].$$

   Compute $R^{\boldsymbol{\lambda}}$ and $R^{\boldsymbol{\lambda}\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda} = \boldsymbol{\lambda}^c$ and let $\Delta(\boldsymbol{\lambda}^c) = -\left[R^{\boldsymbol{\lambda}\boldsymbol{\lambda}}\right]^{-1} R^{\boldsymbol{\lambda}}$.

   If $\|\Delta(\boldsymbol{\lambda}^c)\| < \epsilon$ stop the algorithm and report $\boldsymbol{\lambda}^c$; otherwise continue.

3. Calculate $\boldsymbol{\delta}^c = \gamma^c \Delta(\boldsymbol{\lambda}^c)$. If $1 + (\boldsymbol{\lambda}^c - \delta^c) g_i(\boldsymbol{\beta}) \leq 0$ for some i, let $\gamma^c = \frac{\gamma^c}{2}$ and go to Step 2.

4. Set $\boldsymbol{\lambda}^{c+1} = \boldsymbol{\lambda}^c - \boldsymbol{\delta}^c$, $c = c + 1$, and $\gamma^{c+1} = (c+1)^{-\frac{1}{2}}$ and go to Step 2. Step 2 will guarantee that $p_i > 0$ and the optimization is carried out in the right direction.

4.2 Algorithm for Optimizing Penalized Empirical Likelihood

Let $\hat{\boldsymbol{\lambda}}(\beta)$ be the estimated value of $\boldsymbol{\lambda}$ for a given $\boldsymbol{\beta}$. We maximize the PEL defined in (3.6) over $\boldsymbol{\beta}$. We use the modified Newton-Raphson algorithm proposed by Fan and Li (2001). Note that the penalty function $p_\delta(|\beta_j|)$ is irregular at the origin and may not have a second derivative at some points. Special care is needed in the application of the Newton-Raphson algorithm. Here too, the penalty function is locally approximated as detailed in Section 2 as proposed by Fan and Li (2001). We assume that the profile empirical log-likelihood function is smooth with respect to $\boldsymbol{\beta}$ so that its first two partial derivatives are continuous. Thus, the first term in the profile empirical log-likelihood can be locally approximated via Taylor's expansion. Therefore, the

maximization problem can be reduced to a quadratic maximization problem, and the Newton-Raphson algorithm can be used. The modified Newton-Raphson algorithm for estimating $\boldsymbol{\beta}$ uses quadratic approximation of the profile empirical log-likelihood function. An algorithm for optimizing the penalized empirical likelihood, similar to that in Fan and Li (2001), is as follows:

1. Set $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, and $\epsilon = 1e-08$.

2. Let $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\beta)$ be the estimated value of $\boldsymbol{\lambda}$.

3. The parameter $\boldsymbol{\beta}$ is computed iteratively and the solution at the $(k+1)^{th}$ iteration is given by

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left\{W^{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^k) + n\Sigma_\delta(\boldsymbol{\beta}^k)\right\}^{-1}\left\{W^{\boldsymbol{\beta}}(\boldsymbol{\beta}^k) + nU_\delta(\boldsymbol{\beta}^k)\right\} \tag{4.2}$$

where $W(\boldsymbol{\beta})$ is the profile empirical log-likelihood ratio function defined in (3.5),

$$W^{\boldsymbol{\beta}} = \frac{\partial W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}},\ W^{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{\partial^2 W(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T},$$
$$\Sigma_\delta(\boldsymbol{\beta}^k) = \text{diag}\left[\frac{p'_\delta(|\beta_1^k|)}{|\beta_1^k|}, \ldots, \frac{p'_\delta(|\beta_p^k|)}{|\beta_p^k|}\right], \text{ and } U_\delta(\boldsymbol{\beta}^k) = \Sigma_\delta(\boldsymbol{\beta}^k)\boldsymbol{\beta}^k.$$

Note that to compute $W^{\boldsymbol{\beta}}$ and $W^{\boldsymbol{\beta}\boldsymbol{\beta}}$, we need to estimate the Lagrange multiplier $\hat{\boldsymbol{\lambda}}(\beta)$ as per Section 4.1.

4. If $\min\left|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}\right| < \epsilon$ stop the algorithm and report $\boldsymbol{\beta}^{(k+1)}$; otherwise $k = k+1$ and go to step 3.

We examine the simplified expressions for W^{β} and $W^{\beta\beta}$ as follows. Let R^{β}, $R^{\beta\beta}$, and $R^{\beta\lambda}$ be the first and second partial derivatives of (4.1) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$

$$R^{\beta} = \sum_{i=1}^{n} \left[\frac{g_i'(\boldsymbol{\beta})\boldsymbol{\lambda}}{\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\}} \right], \; R^{\beta\beta} = \sum_{i=1}^{n} \left\{ \left[\frac{g_i''(\boldsymbol{\beta})\boldsymbol{\lambda}^T}{\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\}} \right] - \left[\frac{g_i'(\boldsymbol{\beta})\boldsymbol{\lambda}\boldsymbol{\lambda}^T [g_i'(\boldsymbol{\beta})]^T}{\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\}^2} \right] \right\},$$

and

$$R^{\beta\lambda} = \sum_{i=1}^{n} \left[\frac{\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\} g_i'(\boldsymbol{\beta}) - g_i'(\boldsymbol{\beta})\boldsymbol{\lambda}[g_i(\boldsymbol{\beta})]^T}{\{1 + \boldsymbol{\lambda}^T g_i(\boldsymbol{\beta})\}^2} \right].$$

Now the first derivative of $W(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned} W^{\beta} &= \sum_{i=1}^{n} \left[\frac{\left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}} \right]^T g_i(\boldsymbol{\beta}) + g_i'(\boldsymbol{\beta})\boldsymbol{\lambda}(\beta)}{\{1 + \boldsymbol{\lambda}^T(\beta) g_i(\boldsymbol{\beta})\}} \right] \\ &= \left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}} \right]^T R^{\lambda} + R^{\beta}. \end{aligned}$$

Note that for $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}(\beta)$, $R^{\lambda} = 0$. Therefore,

$$W^{\beta} = R^{\beta}. \tag{4.3}$$

Similarly, the second derivative of $W(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$\begin{aligned} W^{\beta\beta} = \sum_{i=1}^{n} & \left[\frac{\{1 + \boldsymbol{\lambda}^T(\beta) g_i(\boldsymbol{\beta})\} \left\{ \left[\frac{\partial^2 \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] [g_i(\boldsymbol{\beta})]^T + 2 g_i'(\boldsymbol{\beta}) \left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}} \right]^T + g_i''(\boldsymbol{\beta})\boldsymbol{\lambda}(\beta)^T \right\}}{\{1 + \boldsymbol{\lambda}^T(\beta) g_i(\boldsymbol{\beta})\}^2} \right] \\ & - \sum_{i=1}^{n} \left[\frac{\left\{ \left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}} \right]^T g_i(\boldsymbol{\beta}) + g_i'(\boldsymbol{\beta})\boldsymbol{\lambda}(\beta) \right\} \left\{ \left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}} \right]^T g_i(\boldsymbol{\beta}) + g_i'(\boldsymbol{\beta})\boldsymbol{\lambda}(\beta) \right\}^T}{\{1 + \boldsymbol{\lambda}^T(\beta) g_i(\boldsymbol{\beta})\}^2} \right] \end{aligned}$$

$$= \left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}}\right]^T R^{\lambda\lambda} \left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}}\right] + 2\left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}}\right]^T R^{\beta\lambda} + R^{\beta\beta}.$$

Following Owen (2001), a local quadratic approximation to R leads to

$$\left[\frac{\partial \boldsymbol{\lambda}(\beta)}{\partial \boldsymbol{\beta}}\right] = \left(R^{\lambda\lambda}\right)^{-1} R^{\beta\lambda},$$

so that

$$W^{\beta\beta} = R^{\beta\beta} - R^{\beta\lambda}\left(R^{\lambda\lambda}\right)^{-1} R^{\lambda\beta}. \tag{4.4}$$

Optimization over $\boldsymbol{\beta}$ is easier if $W^{\beta\beta}$ is negative definite. The second term in (4.4) is negative semidefinite, but the first term $R^{\beta\beta}$ might not be.

4.3 Selection of Thresholding Parameters

The SCAD penalty function involves two unknown parameters, δ and a. In practice, we could search for the best pair (δ, a) over a two-dimensional structure using cross-validation (CV; Stone, 1974) or generalized cross-validation (GCV; Craven and Wahba, 1979). However, this is computationally expensive. From the Bayesian point of view, Fan and Li (2001) suggested using $a = 3.7$, and this value will be used throughout our simulation studies. Let the empirical likelihood ratio function evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}(\hat{\beta})$ be

$$W(\hat{\boldsymbol{\beta}}) = \left\{ \sum_{i=1}^{n} \log(1 + \hat{\boldsymbol{\lambda}}(\hat{\beta})^T g_i(\hat{\boldsymbol{\beta}}) \right\}.$$

Then, we define the GCV criterion to be

$$\text{GCV}(\delta) = \frac{W(\hat{\boldsymbol{\beta}})}{n\left[1 - e(\delta)/n\right]^2}, \tag{4.5}$$

where $e(\delta)$ is the effective number of regression coefficients given by

$$e(\delta) = \text{tr}\left\{ \left[W^{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) + \Sigma_\delta(\hat{\boldsymbol{\beta}}) \right]^{-1} W^{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) \right\},$$

where $W^{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}})$ is the second derivative of the profile empirical likelihood function with respect to $\boldsymbol{\beta}$ (see (4.4)) evaluated at $\hat{\boldsymbol{\beta}}$, tr denotes the trace of a matrix. We choose the tuning parameters δ to minimize $\text{GCV}(\delta)$.

4.4 Standard Error Formula

The standard errors for the estimated regression parameters can be estimated directly because we are estimating the parameters and selecting the 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 matrix of the estimates $\hat{\boldsymbol{\beta}}$:

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left[\Delta^2 W(\hat{\boldsymbol{\beta}}) + n\Sigma_\delta(\hat{\boldsymbol{\beta}}) \right]^{-1} \widehat{\text{cov}} \left\{ \Delta W(\hat{\boldsymbol{\beta}}) \right\} \left[\Delta^2 W(\hat{\boldsymbol{\beta}}) + n\Sigma_\delta(\hat{\boldsymbol{\beta}}) \right]^{-1}.$$

The covariance matrix of the estimates can be simplified to

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left[\Delta^2 W(\hat{\boldsymbol{\beta}}) + n\Sigma_\delta(\hat{\boldsymbol{\beta}})\right]^{-1} \left[\sum_{i=1}^{n} (n\hat{p}_i)^2 g_i(\hat{\boldsymbol{\beta}}) g_i^T(\hat{\boldsymbol{\beta}})\right] \left[\Delta^2 W(\hat{\boldsymbol{\beta}}) + n\Sigma_\delta(\hat{\boldsymbol{\beta}})\right]^{-1}.$$

Chapter 5

Simulation Studies

We conducted a performance analysis based on a series of Monte-Carlo simulations in linear regression, Poisson regression, and logistic regression and also applied our method to a real-data example. In the simulation studies we compare our method with the penalized-likelihood SCAD method. Our performance measures for these comparisons are the median of the relative model error (MRME), the average number of estimated zero coefficients that are initially set to zero, and the average number of zero coefficients that are not initially set to zero. We also compare the estimated values of the nonzero coefficients and the corresponding standard errors.

Median Relative Model Error (MRME)

Following Tibshirani (1996), we compare the median of the relative model error (Fan and Li, 2001) rather than the mean relative model error because of the instability of the best-subset variable selection. The model error for the linear model is defined by

$$ME(\hat{\boldsymbol{\beta}}) = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^T E(X^T X) \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

The error for the selected model is compared to the error of the full model. For each variable selection method, we computed the median of the relative model error, and this is reported in the simulation studies.

5.1 Linear Regression Model

Consider a linear model of the form

$$y_i = \boldsymbol{X}_i \boldsymbol{\beta} + \sigma \epsilon_i \tag{5.1}$$

with $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ where $\boldsymbol{X}_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_p})$ is a vector of covariates and $p = 8$. The components of $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ are standard normal, the correlation between x_i and x_j is $0.5^{|i-j|}$, and σ=1. The least-squares estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{LS} = \left[\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{X}_i\right]^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^T y_i = \left[\boldsymbol{X}^T \boldsymbol{X}\right]^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{5.2}$$

The estimating equation for $\boldsymbol{\beta}$ is given by

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i^T [y_i - \boldsymbol{X}_i \boldsymbol{\beta}] = 0 \tag{5.3}$$

and the first derivative of the estimating equation $g(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$g'(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{X}_i.$$

We simulated 10000 data sets with $n = 60$ observations from the above model with the components of $\boldsymbol{X}$ and $\boldsymbol{\epsilon}$ being standard normal. This is the model used by Tibshirani (1996). Our penalized-empirical-likelihood SCAD (PELSCAD) is compared only with SCAD since Fan and Li (2001) reported that SCAD performs better than LASSO and other information-theoretic approaches. Following Tibshirani (1996) and Fan and Li (2001), the performance of these methods was assessed based on MRME and the number of zero coefficients. We also repeated the entire study with sample size $n = 100$. The MRME values based on 10000 simulated data sets are summarized in Table 5.1. It also reports the average number of zero and nonzero coefficients. The column labeled "Correct" gives the average number of estimated zero coefficients that were initially set to zero, and the column labeled "Incorrect" gives the average number of zero coefficients that were not initially set to zero. The estimated values of the nonzero coefficients and the corresponding standard errors are reported in Table 5.2. From Table 5.1 we see that for $n = 60$ the MRME of SCAD is slightly smaller

than that of PELSCAD, and for both methods the average number of zero coefficients is close to the target of five. When the sample size increases to 100, the MRME of PELSCAD is low compared to that of SCAD. The average number of zero coefficients is again close to five. This clearly indicates that both methods perform well when a parametric model is available.

Method	MRME%	Avg. no. of zero coefficients	
		Correct	Incorrect
n=60, σ=1			
SCAD	35.57	4.61	0.0
PELSCAD	36.52	4.61	0.0
n=100, σ=1			
SCAD	41.50	4.85	0.0
PELSCAD	34.55	4.95	0.0

Table 5.1: Simulation results for linear regression model

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$
n=60, σ=1			
SCAD	3.015	1.474	2.003
	(0.167)	(0.195)	(0.136)
PELSCAD	3.002	1.496	1.999
	(0.163)	(0.170)	(0.141)
n=100, σ=1			
SCAD	3.027	1.442	2.003
	(0.139)	(0.185)	(0.104)
PELSCAD	2.999	1.499	1.999
	(0.120)	(0.124)	(0.104)

Table 5.2: Linear regression model: Estimates of nonzero coefficients with corresponding standard errors in parentheses

5.2 Poisson Regression Model

In this section, we consider the performance of our method when the parametric model is misspecified, in the context of a Poisson regression model. Let $y_1, y_2, ..., y_n$ be n independent responses, each of which follows a Poisson distribution. The relationship between the mean and variance is given by $\mathrm{E}(y_i) = \mu_i = \mathrm{Var}(y_i)$.

Let $\boldsymbol{\mu}^T = (\mu_1, \mu_2, ..., \mu_n)$. Let $\boldsymbol{X}$ be the design matrix and assume that the components of $\boldsymbol{X}$ are standard normal. Assume also that

$$log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$$

where $\boldsymbol{\beta} \in \mathcal{R}^p$ is the vector of regression coefficients. Then, the log likelihood function for $\boldsymbol{\beta}$ is given by

$$l[\boldsymbol{\beta}; \boldsymbol{y}] = \sum_{i=1}^{n} \{y_i \boldsymbol{X}_i \boldsymbol{\beta} - \exp(\boldsymbol{X}_i \boldsymbol{\beta})\}. \tag{5.4}$$

The estimating equation for $\boldsymbol{\beta}$ is given by

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i^T (y_i - \exp(\boldsymbol{X}_i \boldsymbol{\beta}))$$

and the first derivative of the estimating equation $g(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$g'(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \exp(\boldsymbol{X}_i \boldsymbol{\beta}) \boldsymbol{X}_i^T \boldsymbol{X}_i.$$

We generate over-dispersed Poisson count data $\boldsymbol{y}$ using the model specified through a conditional density given by

$$f(y_i|u_i,\mu) = \frac{(u_i\mu)^{y_i}\ \exp(-u_i\mu)}{y_i!}, \quad i = 1, 2, ..., n \tag{5.5}$$

with u_i a random variable such that $\mathrm{E}(u_i) = 1$ and $\mathrm{Var}(u_i) = \omega$. Marginally, we have $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{\mu}(1 + \boldsymbol{\mu}\omega)$. The distribution of u is chosen to be gamma with parameters $(\omega, 1/\omega)$ with ω being the over-dispersion parameter. However, the parametric likelihood and empirical likelihood are constructed under the assumption that there is no over-dispersion. We consider a four-covariate generalized linear model such that

$$log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

with $\beta = (0.5, 0.5, 0.6, 0, 0)$. The covariates $\mathbf{X} = (x_1, x_2, x_3, x_4)$ are generated from a multivariate normal distribution with mean zero, and the correlation between x_i and x_j is $0.5^{|i-j|}$. We choose four levels of over-dispersion: $\omega = 0, 1/8, 1/6, 1/4$. Note that when $\omega = 0$, we use ordinary Poisson regression model to generate the responses. Where as for $\omega > 0$, we use the conditional density model (5.5) to generate the responses. This is the simulation model used by Variyath, Chen, and Abraham (2010). In each simulation, we generate $n = 100$ observations for the response $\boldsymbol{y}$ from the conditional distribution specified earlier. For each model, we analyze 10000 simulated

data sets. The MRME and the average number of zero and nonzero coefficients over 10000 simulated data sets are summarized in Table 5.3. The estimated values of the nonzero coefficients and the corresponding standard errors are reported in Table 5.4. From Table 5.3 we see that when there is no over-dispersion ($\omega = 0$), the MRME of PELSCAD is smaller than that of SCAD. The average number of zero coefficients for PELSCAD is closer to the target of two in all cases. When the over-dispersion increases, PELSCAD performs better than SCAD. From Table 5.4 we see that the nonzero parameter estimates of PELSCAD are close to the true values and the SCAD estimates are not as close. Note that in PELSCAD, we did not model the over-dispersion.

Method	MRME%	Avg. no. of zero coefficients	
		Correct	Incorrect
n=100, ω=0			
SCAD	79.42	1.41	0.0004
PELSCAD	58.49	1.74	0.0001
n=100, ω=1/8			
SCAD	86.24	1.24	0.0010
PELSCAD	68.90	1.61	0.0003
n=100, ω=1/6			
SCAD	89.91	1.19	0.0012
PELSCAD	65.86	1.64	0.0005
n=100, ω=1/4			
SCAD	88.61	1.12	0.0028
PELSCAD	69.95	1.62	0.0033

Table 5.3: Simulation results for Poisson regression model

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
n=100, ω=0			
SCAD	0.455	0.502	0.611
	(0.113)	(0.089)	(0.087)
PELSCAD	0.515	0.498	0.601
	(0.089)	(0.078)	(0.088)
n=100, ω=1/8			
SCAD	0.450	0.492	0.602
	(0.127)	(0.115)	(0.115)
PELSCAD	0.502	0.495	0.589
	(0.106)	(0.106)	(0.108)
n=100, ω=1/6			
SCAD	0.448	0.488	0.601
	(0.134)	(0.123)	(0.122)
PELSCAD	0.506	0.497	0.587
	(0.107)	(0.107)	(0.113)
n=100, ω=1/4			
SCAD	0.444	0.483	0.597
	(0.139)	(0.135)	(0.134)
PELSCAD	0.482	0.495	0.597
	(0.124)	(0.127)	(0.129)

Table 5.4: Poisson regression: Estimates of nonzero coefficients with corresponding standard errors in parentheses

5.3 Logistic Regression Model

Let $y_1, y_2, ..., y_n$ be n independent Bernoulli trials with mean and variance $E(y_i) = \pi_i$ and $\text{Var}(y_i) = \pi_i(1 - \pi_i)$, where $\boldsymbol{\pi}^T = (\pi_1, \pi_2, ..., \pi_n)$. Let $\boldsymbol{X}$ be the design matrix and $\boldsymbol{\beta}$ a $p \times 1$ vector of regression coefficients. Assume that

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \boldsymbol{X}_i\boldsymbol{\beta}.$$

The log-likelihood function for $\boldsymbol{\beta}$ is

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \{y_i \boldsymbol{X}_i\boldsymbol{\beta} - \log\left[1 + \exp(\boldsymbol{X}_i\boldsymbol{\beta})\right]\}. \tag{5.6}$$

The estimating function for $\boldsymbol{\beta}$ can be written

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i^T(y_i - \pi_i),$$

where

$$\pi_i = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{\{1 + \exp(\boldsymbol{X}_i\boldsymbol{\beta})\}}.$$

The first derivative of the estimating function $g(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is

$$g'(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \pi_i(1 - \pi_i)\boldsymbol{X}_i^T\boldsymbol{X}.$$

We generate n=200 observations for the response $\boldsymbol{y}$ from the model

$$y_i \sim \text{Bernoulli}\{p(\boldsymbol{X}_i\boldsymbol{\beta})\},$$

where

$$p(\boldsymbol{X}_i\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{X}_i\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i\boldsymbol{\beta})},$$

and the first six components of $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are as for the linear regression model discussed in Section 5.1. The last two components of $\boldsymbol{X}$ are assumed to have a Bernoulli distribution with probability of success 0.5. All covariates are standardized. We repeat the simulation studies for $n = 500$ and $n = 1000$. Fan and Li (2001) used a similar logistic regression model for comparison purposes. The simulation results are summarized in Tables 5.5 and 5.6. From Table 5.5 we see that PELSCAD has a smaller MRME than SCAD for all sample sizes. If the sample size is increased, the MRMEs are closer to each other and the average number of zero coefficients is also closer to the target value of five. Overall, the PELSCAD method performs well in this case too.

5.4 Australian Health Survey

We consider the data set for doctor visits from the Australian health survey of 1977–78. It contains health information for 5190 single adults where the young and old have been oversampled. The data set is also available in the "R" statistical software (in the faraway library). We apply variable selection methods under Poisson regression

Method	MRME%	Avg. no. of zero coefficients	
		Correct	Incorrect
n=200, σ=1			
SCAD	67.01	4.83	0.0110
PELSCAD	54.25	4.86	0.0031
n=500, σ=1			
SCAD	57.07	4.99	0.0004
PELSCAD	56.41	4.84	0.0000
n=1000, σ=1			
SCAD	55.86	5.00	0.0000
PELSCAD	53.33	4.98	0.0000

Table 5.5: Simulation results for logistic regression model

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$
n=200, σ=1			
SCAD	3.450	1.705	2.301
	(0.660)	(0.503)	(0.493)
PELSCAD	3.276	1.662	2.186
	(0.703)	(0.455)	(0.520)
n=500, σ=1			
SCAD	3.211	1.605	2.138
	(0.331)	(0.249)	(0.249)
PELSCAD	3.087	1.55	2.06
	(0.355)	(0.246)	(0.263)
n=1000, σ=1			
SCAD	3.132	1.568	2.088
	(0.224)	(0.164)	(0.166)
PELSCAD	3.034	1.519	2.024
	(0.238)	(0.168)	(0.175)

Table 5.6: Logistic regression model: Estimates of nonzero coefficients with corresponding standard errors in parentheses

to this data set. The response of interest is the health of adults, which is measured in terms of the number of consultations with a doctor or specialist in the previous two weeks ($\boldsymbol{y}$). In addition, we have several measures of health service utilization and socio-economic parameters. Cameron et al. (1988) analyzed this data set using an economic model of the joint determination of health service use and health-insurance choices in Australia. Cameron and Trivedi (1986) studied this data set in a different context. Our main objectives are to model the relationship between the response and the covariates and to identify the simplest model that gives a clear picture of the data-generating structure. A short description of the variables is given in Table 1.1 of Chapter 1.

The mean of the response, number of doctor visits, is 0.302 and the standard deviation is 0.798. The data indicate that there is over-dispersion. The estimates of the Poisson regression coefficients are given in Table 5.7. From this table, we see that illness (X_8) and actdays (X_9) are statistically significant. The covariate sex (X_1) is also marginally significant, indicating that female patients visit doctors more frequently than male patients do. We use penalized-empirical-likelihood SCAD (PELSCAD) and parametric SCAD to select the significant covariates for this real-data example. We compare the results with information-theoretic approaches such as AIC and BIC

in their empirical-likelihood versions. The selected covariates, the corresponding estimates of the regression parameters, and their standard errors are listed in Table 5.8 for each method. From this table, we see that SCAD and PELSCAD identified the covariates age (X_2), illness (X_8), actdays (X_9), and hscore (X_{10}) as important and forced regression coefficients of the other variables to zero. Note that the empirical-likelihood version of BIC (EBIC) selected the simplest model whereas AIC selected the largest model. These results are useful for understanding the data-generating mechanism and for prediction.

Variables	Coefficients	Standard Error	z-value	$P[Z > z]$
y-Dvisits	-2.2238	0.1898	-11.716	<10e-16
X_1-Sex	0.1569	0.0561	2.795	0.0052
X_2-Age	1.0563	1.0007	1.055	0.2912
X_3-Agesq	-0.8487	1.0778	-0.787	0.4310
X_4-Income	-0.2053	0.0884	-2.323	0.0202
X_5-Levyplus	0.1232	0.0716	1.720	0.0855
X_6-Freepoor	-0.4401	0.1798	-2.447	0.0144
X_7-Freerepa	0.0798	0.0921	0.867	0.3861
X_8-Illness	0.1869	0.0183	10.227	<10e-16
X_9-Actdays	0.1268	0.0050	25.198	<10e-16
X_{10}-Hscore	0.0301	0.0101	2.979	0.0029
X_{11}-Chcond1	0.1141	0.0666	1.712	0.0869
X_{12}-Chcond2	0.1412	0.0831	1.698	0.0896

Table 5.7: Estimates of Poisson regression coefficients for full model

Variable	AIC	EAIC	BIC	EBIC	SCAD	PELSCAD
Intercept	-2.0891	-2.2049	-2.2444	-2.0486	-2.010	-1.9952
	(0.1008)	(0.0691)	(0.0679)	(0.0517)	(0.0626)	(0.0191)
X_1	0.1620	0.2003	0.2056	0.2627	—	—
	(0.0558)	(0.0542)	(0.0542)	(0.0527)	—	—
X_2	0.3551	0.5168	0.5694	—	0.9970	1.1507
	(0.1432)	(0.1319)	(0.1307)	—	(0.1231)	(0.0462)
X_4	-0.1998	—	—	—	—	—
	(0.0843)	—	—	—	—	—
X_5	0.0837	—	—	—	—	—
	(0.0535)	—	—	—	—	—
X_6	-0.4696	-0.4375	—	—	—	—
	(0.1764)	(0.1731)	—	—	—	—
X_8	1.1861	0.1988	0.1997	0.2303	0.0638	0.0204
	(0.0183)	(0.0175)	(0.0175)	(0.0165)	(0.0044)	(0.0004)
X_9	0.1266	0.1277	0.1279	0.1363	0.1299	0.1371
	(0.0050)	(0.0049)	(0.0049)	(0.0045)	(0.0041)	(0.0033)
X_{10}	0.0311	0.0334	0.0320	—	0.0127	0.0047
	(0.0050)	(0.0049)	(0.0049)	—	(0.0016)	(0.0003)
X_{11}	0.1211	—	—	—	—	—
	(0.0664)	—	—	—	—	—
X_{12}	0.1589	—	—	—	—	—
	(0.0818)	—	—	—	—	—

Table 5.8: Estimates of Poisson regression coefficients, with their standard errors in parentheses, for model identified by different variable selection methods

Chapter 6

Variable Selection for Cox's Proportional Hazard Model

Variable selection is an important problem in survival analysis. In practice, many covariates are potential risk factors and at the initial stage of the modeling, we normally introduce a large number of predictors. Thus, the selection of significant risk factors plays a crucial role in survival analysis. We focus our attention on Cox's proportional hazards model with right-censored survival data (Lindley, 1968).

Bayesian model-selection procedures for survival analysis have been proposed by Faraggi and Simon (1997) and Faraggi (1998). Ibrahim, Chen, and MacEachern (1999) proposed a full Bayesian variable selection procedure for the Cox model by

specifying a nonparametric prior for the baseline function and a parametric prior for the regression coefficients. Bayesian variable selection procedures are simple, but hard to implement especially in high-dimensional modeling because of the computational burden of the calculation of the posterior model probabilities. Some traditional variable selection criteria such as AIC and BIC can easily be extended to survival analysis. Volinsky and Raftery (2000) extended BIC to the Cox model. Other traditional variable selection procedures such as stepwise deletion and best-subset selection are useful in practice. However, they suffer from several drawbacks, the most severe of which is a lack of stability (for more details see Chapter 1). Tibshirani (1997) extended the LASSO variable selection procedures to the Cox model. Fan and Li (2002) derived a nonconcave penalized partial likelihood for the Cox model and illustrated the oracle properties of their procedures.

In this chapter, we introduce the penalized empirical likelihood for Cox's proportional hazards model. A comprehensive review of empirical likelihood has been given in Chapter 3. There are many recent studies of EL for survival analysis. Empirical likelihood has many nice properties, including the ability to carry out hypothesis tests and construct confidence intervals without estimating the variance. This is possible because the EL ratio does not involve the unknown variances and the limiting distribution of EL is chi square. This feature has been useful in survival analysis because

variance estimation can be difficult in these problems. In EL, we need not estimate the variances which makes many inference procedures practical. We now introduce in detail the survival function, hazard function, and right-censored data of survival analysis.

Survival Function

Let T be a nonnegative random variable with distribution function f representing the failure time of an individual from a homogeneous population. The survival function is a tool to describe time-to-event phenomena. It captures the probability of an individual surviving beyond a specific time t. It is defined as

$$\begin{aligned} S(t) = Pr(T \geq t) &= 1 - Pr(T < t) \\ &= 1 - F(t). \end{aligned}$$

$S(t)$ is referred to as the reliability function in the context of failure time. If T is a continuous random variable, then $S(t)$ is a continuous and monotonically decreasing function. The survival function can be written

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

Thus,

$$f(t) = -\frac{dS(t)}{dt}.$$

Hazard Function

The hazard function is a fundamental quantity in survival analysis. This function is known as the conditional failure rate in reliability. The hazard rate is defined by

$$h(t) = \lim_{\Delta t \to 0} \frac{P\left[t \leq T \leq t + \Delta t | T \geq t\right]}{\Delta t}.$$

If T is a continuous random variable, then we can show that

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d \ln[S(x)]}{dx}.$$

A related quantity is the cumulative or integrated hazard function $H(t)$, defined by

$$H(t) = \int_0^t h(u)du = -\ln[S(t)].$$

Thus, for continuous lifetimes,

$$S(t) = \exp\left[-H(t)\right] = \exp\left[-\int_0^t h(u)du\right].$$

Right Censoring

In survival analysis, censoring refers to data that are missing for some random reason. If the birth and death dates of a individual are known, then the lifetime is known. However, we may know only that the date of death is after for some date; this is called right censoring. Right censoring occurs for those individuals whose birth date is known but who are still alive when they are lost to follow-up or when the study

ends. These censoring times may vary from individual to individual.

In right censoring, for a specific individual under study, we assume that there is a lifetime T and a right-censoring time C. The T's are random variables with density function $f(t)$ and survival function $S(t)$. The exact lifetime T of an individual is less than or equal to C. If T is greater than C, then the individual is a survivor, and his or her event time is censored at C. The survival data can be conveniently represented by pairs of random variables (T, δ), where δ indicates whether the lifetime T corresponds to an event ($\delta = 1$) or is censored ($\delta = 0$), and Z is equal to T if the lifetime is observed, and to C if it is censored, i.e., $Z = \min\{T, C\}$.

We construct the likelihood function for right censoring as follows. For $\delta = 0$,

$$\begin{aligned} P[Z, \delta = 0] &= Pr[Z = C|\delta = 0]Pr[\delta = 0] = Pr[\delta = 0] \\ &= Pr[Z > C] = S(C). \end{aligned}$$

For $\delta = 1$,

$$\begin{aligned} P[Z, \delta = 1] &= Pr[Z = T|\delta = 1]Pr[\delta = 1] \\ &= Pr[Z = T|T \leq C]Pr[T \leq C] \\ &= \left[\frac{f(z)}{1 - S(C)}\right][1 - S(C)] = f(z). \end{aligned}$$

This can be combined into a single expression,

$$Pr[z, \delta] = [f(t)]^{\delta}\,[S(z)]^{1-\delta}.$$

For a random sample of pairs $(Z_i, \delta_i), i = 1, \ldots, n$, the likelihood function can be written

$$L = \prod_{i=1}^{n} Pr[z_i, \delta_i] = \prod_{i=1}^{n} [f(z_i)]^{\delta_i} [S(z_i)]^{1-\delta_i}.$$

6.1 Proportional Hazards Model

Let $\boldsymbol{T}, \boldsymbol{C}$, and $\boldsymbol{X}$ be respectively the survival time, the censoring time, and the associated covariate values. Let $\boldsymbol{Z} = \min\{\boldsymbol{T}, \boldsymbol{C}\}$ be the observed time and $\boldsymbol{\delta} = I(\boldsymbol{T} \leq \boldsymbol{C})$ be the event indicator ($\delta = 1$ if the event has occurred and $\delta = 0$ if the lifetime is right-censored). We assume that $\boldsymbol{T}$ and $\boldsymbol{C}$ are conditionally independent given $\boldsymbol{X}$ and the censoring system is noninformative.

Our observed data $\{(\boldsymbol{X}_i, Z_i, \delta_i) : i = 1, \ldots, n\}$ are a random sample from a certain population $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\delta})$. The complete likelihood of the data is given by

$$\begin{aligned} L &= \prod_{i=1}^{n} [f(Z_i|\boldsymbol{X}_i)]^{\delta_i} [S(Z_i|\boldsymbol{X}_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{f(Z_i|\boldsymbol{X}_i)}{S(Z_i|\boldsymbol{X}_i)}\right]^{\delta_i} [S(Z_i|\boldsymbol{X}_i)] \\ &= \prod_{i=1}^{n} [h(Z_i|\boldsymbol{X}_i)]^{\delta_i} [S(Z_i|\boldsymbol{X}_i)]. \end{aligned}$$

The complete likelihood simplifies to

$$L = \prod_{i=1}^{n} [h(Z_i|\boldsymbol{X}_i)]^{\delta_i} \prod_{i=1}^{n} \exp\{-H(Z_i|\boldsymbol{X}_i)\}. \tag{6.1}$$

To present this likelihood function clearly for Cox's proportional hazards model, we need more notation. Let $\tau_1 < \tau_2 \ldots < \tau_N$ denote the ordered observed failure time corresponding to $t_1, t_2, \ldots, t_n$. Let (j) be the label for the item failing at τ_j, and let the covariates associated with N failures be $\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \ldots, \boldsymbol{X}_{(N)}$. Let R_j denote the risk set immediately before time τ_j, defined by

$$R_j = \{i : Z_i \geq \tau_j\}$$

Consider the proportional hazards model proposed by Cox (1975):

$$h(t|\boldsymbol{X}) = h_0(t) \exp(\boldsymbol{X}\boldsymbol{\beta}),$$

where $\boldsymbol{X}_i^T$ is a $p \times 1$ vector of covariates, $\boldsymbol{\beta}$ is $p \times 1$ vector of parameters, and $h_0(t)$ is the baseline hazard function. The likelihood in (6.1) becomes

$$L = \prod_{i=1}^{N} [h_0(Z_{(i)}) \exp(\boldsymbol{X}_{(i)}\boldsymbol{\beta})] \prod_{i=1}^{n} \exp\{-H_0(z_i) \exp(\boldsymbol{X}_i\boldsymbol{\beta})\} \tag{6.2}$$

where $H_0(*)$ is the cumulative baseline hazard function. In the Cox proportional hazards model, the baseline hazard function is unknown and is not parameterized. Following the idea of Breslow (1975), consider the "least informative" nonparametric model for $H_0(*)$, in which $H_0(t)$ has a possible jump h_j at the observed failure time τ_j. More precisely, let

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

Then

$$H_0(Z_i) = \sum_{j=1}^{N} h_j I(i \in R_j). \tag{6.3}$$

Using (6.3), the log-likelihood of (6.2) becomes

$$\ell(h_0(\boldsymbol{Z})) = \sum_{j=1}^{N} \left\{ \log(h_j) + \boldsymbol{X}_{(j)}\boldsymbol{\beta} \right\} - \sum_{i=1}^{n} \left\{ \sum_{j=1}^{N} h_j I(i \in R_j) \exp(\boldsymbol{X}_i\boldsymbol{\beta}) \right\}. \tag{6.4}$$

Taking the partial derivative of $\ell(h_0(\boldsymbol{Z}))$ with respect to h_j and equating to zero gives

$$\hat{h}_j = \frac{1}{\left\{ \sum_{i \in R_j} \exp(\boldsymbol{X}_i\boldsymbol{\beta}) \right\}}.$$

Substituting $\hat{h}_j$ into (6.4) and removing the constant term $-N$, we can write the partial log-likelihood as

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{N} \left[\boldsymbol{X}_{(j)}\boldsymbol{\beta} - \log \left\{ \sum_{i \in R_j} \exp(\boldsymbol{X}_i\boldsymbol{\beta}) \right\} \right].$$

An equivalent way of writing the partial log-likelihood is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{X}_i\boldsymbol{\beta} - \log \left\{ \sum_{l}^{n} \exp(\boldsymbol{X}_l\boldsymbol{\beta}) Y_l(Z_i) \right\} \right] \tag{6.5}$$

where $Y_l(u) = I(Z_i \geq u)$ indicates whether or not the i^{th} individual is at risk at time u. Taking the partial derivative of (6.5) with respect to $\boldsymbol{\beta}$ and equating to zero gives

the estimating equation for $\boldsymbol{\beta}$. This can be written

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{X}_i^T - \frac{\sum_{l}^{n} \exp(\boldsymbol{X}_l \boldsymbol{\beta}) Y_l(Z_i) \boldsymbol{X}_l^T}{\sum_{l}^{n} \exp(\boldsymbol{X}_l \boldsymbol{\beta}) Y_l(Z_i)} \right]. \tag{6.6}$$

The empirical likelihood method has been extended to linear regression with censored data (Qin and Jing, 2001a; Li and Wang, 2003; Qin and Tsao, 2003). It has also been adapted for semiparametric regression models, including partial linear models (Leblanc and Crowley, 1995; Shen, Shi, and Wong, 1999; Qin and Jing, 2001b; Lu, Chen, and Gan, 2002; Wang and Li, 2002). We propose a nonparametric version of the penalized-likelihood variable selection method in survival analysis, replacing the parametric likelihood by the empirical likelihood. Following equation (3.6), we can write the penalized empirical log-likelihood function for Cox's proportional hazard model as

$$\mathbf{L}(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \left[\log(1 + \hat{\boldsymbol{\lambda}}^T(\boldsymbol{\beta}) g_i(\boldsymbol{\beta})) \right] - n \sum_{j=1}^{p} p_\delta(|\beta_j|) \tag{6.7}$$

where $p_\delta(*)$ is the SCAD penalty function defined in (2.2) and $g(\boldsymbol{\beta})$ is defined in (6.6). The penalized empirical likelihood estimate of $\boldsymbol{\beta}$ is derived by maximizing (6.7) with respect to $\boldsymbol{\beta}$, with the proper choice of the tuning parameters involved in the SCAD penalty function. For the maximization, we used the modified Newton-Raphson algorithm discussed in Chapter 4. During the maximization, many of the

insignificant estimated coefficients are forced to zero and hence their corresponding variables do not appear in the model. This achieves the objective of the variable selection.

6.2 Simulation Studies

Fan and Li (2002) conducted a series of Monte-Carlo simulations for Cox's proportional hazards model and showed that the penalized-likelihood variable selection using SCAD has better performance than the LASSO, HARD, best-subset, and Oracle variable selection methods. Consider the exponential hazard model

$$h(t|\boldsymbol{X}) = \exp(\boldsymbol{X}\boldsymbol{\beta}),$$

with $\boldsymbol{\beta}_0 = (0.8, 0, 0, 1, 0, 0, 0.6, 0)^T$. Let the correlation between x_i and x_j be $\rho^{|i-j|}$. The distribution of the censoring time is exponential with mean $U\exp(\boldsymbol{X}\boldsymbol{\beta}_0)$, where U is randomly generated from the uniform distribution over $[1, 3]$ for each simulated data set, so that about 30% of the data are censored. We simulated 1000 data sets consisting of n =75 and 100 and $\rho = 0.3$ and 0.5 from the exponential hazard model with the components of $\boldsymbol{X}$ being standard normal. This model is used by Fan and Li (2002). The model errors of our procedures are compared to those of Cox's estimates. The median of the relative model error (MRME) and the average number

of zero coefficients over 1000 simulated data sets are summarized in Table 6.1. The estimated values of the nonzero coefficients and the corresponding standard errors are reported in Table 6.2.

From Table 6.1 we see that when $\rho = 0.3$ and $n = 75$ or 100 the MRME of PELSCAD is smaller than that of SCAD and the average number of zero coefficients is closer to the target of five. From Table 6.2 we see that the nonzero parameter estimates of PELSCAD and SCAD are close to the true values and their corresponding standard errors (given in parentheses). Similar results hold for $\rho = 0.5$; see Tables 6.3 and 6.4. This clearly indicates that PELSCAD performs well compared to SCAD.

Method	MRME	Ave. no. of 0 coefficients	
		Correct	Incorrect
n=75, $\rho = 0.3$			
SCAD	44.58	4.43	0.000
PELSCAD	19.77	4.98	0.049
n=100, $\rho = 0.3$			
SCAD	37.21	4.62	0.000
PELSCAD	22.05	5.00	0.017

Table 6.1: Simulation results for Cox's proportional hazards model

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$
n=75, $\rho = 0.3$			
SCAD	0.827	1.042	0.595
	(0.186)	(0.180)	(0.219)
PELSCAD	0.834	1.048	0.616
	(0.178)	(0.185)	(0.226)
n=100, $\rho = 0.3$			
SCAD	0.825	1.029	0.605
	(0.148)	(0.148)	(0.161)
PELSCAD	0.824	1.030	0.611
	(0.144)	(0.145)	(0.181)

Table 6.2: Cox's proportional hazards model: Estimates of nonzero coefficients with corresponding standard errors in parentheses

Method	MRME	Ave. no. of 0 coefficients	
		Correct	Incorrect
n=75, $\rho = 0.5$			
SCAD	41.59	4.56	0.025
PELSCAD	22.58	4.88	0.046
n=100, $\rho = 0.5$			
SCAD	37.51	4.73	0.008
PELSCAD	20.69	4.97	0.014

Table 6.3: Simulation results for Cox's proportional hazards model

Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_5$
n=75, $\rho = 0.5$			
SCAD	0.828	1.042	0.595
	(0.187)	(0.203)	(0.225)
PELSCAD	0.844	1.046	0.616
	(0.184)	(0.197)	(0.229)
n=100, $\rho = 0.5$			
SCAD	0.818	1.028	0.600
	(0.145)	(0.153)	(0.174)
PELSCAD	0.830	1.038	0.620
	(0.142)	(0.164)	(0.165)

Table 6.4: Cox's proportional hazards model: Estimates of nonzero coefficients with corresponding standard errors in parentheses

6.3 Lung Cancer Example

We now apply our variable selection method to the lung-cancer data set. The data set, lung.data, is available in the "R" statistical package (in the SIS library). This data set contains information on 137 subjects, such as survival time and censor status, as well as information on six covariates. The covariates are X_1 =trt (1=standard treatment and 2=test), X_2 = celltype (1=squamous, 2=smallcell, 3=adeno, and 4=large), X_3 = karno (Karnofsky performance score), X_4 = diagtime (months from diagnosis to randomization), X_5 = age (in years), and X_6 =prior (prior therapy: 0=no and 1=yes).

The regression estimates of the full Cox's proportional hazards model are given in

Variables	Coefficients	Standard Error	z-value	$P[Z > z]$
X_1-trt	0.1138	0.0944	1.206	0.2279
X_2-celltype	0.1383	0.0828	1.670	0.0949
X_3-karno	-0.7080	0.1082	-6.541	<10e-11
X_4-diagtime	0.0230	0.0962	0.239	0.8113
X_5-age	-0.0384	0.0965	-0.398	0.6907
X_6-prior	-0.0358	0.1017	-0.352	0.7246

Table 6.5: Estimates of Cox's proportional hazards model coefficients for full model

Table 6.5. From this table, we see that the Karnofsky performance score (X_3) is statistically significant. We now use the penalized-empirical-likelihood SCAD (PELSCAD) and SCAD procedures to select the significant covariates for this real-data example. The selected covariates and the corresponding estimates of the regression parameters with their standard errors in parentheses are listed in Table 6.6. From this table, we see that SCAD and PELSCAD identified only the covariate Karnofsky performance score (X_3) as important; the other variables were not selected in the final model. These results are useful for understanding the data-generating mechanism, for fitting a simple model, and for prediction.

Variables	SCAD	PELSCAD
X_1-trt	—	—
X_2-celltype	—	—
X_3-karno	-0.6713	-0.6672
	(0.0917)	(0.1220)
X_4-diagtime	—	—
X_5-age	—	—
X_6-prior	—	—

Table 6.6: Estimates of regression coefficients in Cox's proportional hazards model

Chapter 7

Conclusion

In this chapter, we summarize our contributions to variable selection. We proposed a penalized variable selection approach based on the empirical likelihood (EL), a nonparametric likelihood sharing many of the properties of the parametric likelihood. In our method we do not need to specify the parametric family of the distribution to do the inference. We defined the profile empirical likelihood based on a set of estimating equations and developed penalized-empirical-likelihood variable selection. We discussed the asymptotic properties of our method in detail. We also proposed an algorithm for the implementation of the new method. Simulation studies showed that our method is consistent and when a parametric model is available its performance is comparable to that of the existing method for linear regression, Poisson regression,

and logistic regression. When the parametric model is misspecified, our method outperforms the existing method. We also applied our method to survival analysis to investigate variable selection in Cox's proportional hazards model. Our simulation showed that PELSCAD performs as well as SCAD.

Bibliography

[1] Akaike, H (1973). Information theory and an extension of the maximum likelihood principle. In Petrove, B. N. and Csaki, F. (eds.) *Second Symposium of Information Theory*, Akademiai Kiado, Budapest, 267-282.

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

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

[4] Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations. *Journal of the American Statistical Association* **96**, 939-967.

[5] Bickel, P. and Freedman, D. (1982). Bootstrapping regression models with many variables: A festchrift for E.L. Lehmann. Belmont, CA: Wadsworth, 28-48.

[6] Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression: The X-random case. *International Statistical Review* ***60****, 291-319.*

[7] Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistics Review* ***43****, 45-58.*

[8] Cameron, A. C. and Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of estimators. *Journal of Applied Econometrics* ***1****, 29-53.*

[9] Cameron, A. C and Trivedi, P. K. (1998). Regression Analysis of Count Data. Cambridge University Press, United Kingdom.

[10] Cameron, A. C., Trivedi, P. K., Milne, F., and Piggot, J. (1988). A microeconometric model of the demand for health care and health insurance in Australia. *Review of Economic Studies* ***55****, 85-106.*

[11] Chen, J. and Qin, J. (1993). Empirical likelihood estimation for fininite populations and the effective usage of auxiliary information. *Biometrika* ***80****, 107-116.*

[12] Chen, J., Sitter, R. R., and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika* ***89****, 230-237.*

[13] Chen, J.,Variyath, A. M. and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational Graphics and statistics* ***17****, 426-443.*

[14] Chen, S. X. (1993). On the accuracy of empirical likelihood confidence regions for linear regression model. *Annals of the Institute of Statistical Mathematics* ***45****, 621-637.*

[15] Chen, S. X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis* ***49****, 24-40.*

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

[17] Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematika* ***31****, 377-403.*

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

[19] Efroymson, M. A. (1960). Multiple regression analysis. In Ralsston, A. and Wilf, H. S. (eds.), *Mathematical Methods for Digital Computers.* Wiley, New York.

[20] Fan, J. (1997). Comments on "Wavelets in statistics: A review" by A. Antoniadis. *Journal of the Italian Statistical Association* ***6****, 131-138.*

[21] Fan, J. and Li, R. (2001). Variable selection via non concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* ***96****, 1348-1360.*

[22] Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics* ***30****, 74-99.*

[23] Faraggi, D. (1998). Bayesian variable selection method for censored survival data. *Biometrika* ***54****, 1475-1485.*

[24] Faraggi, D. and Simon, R. (1997). Large sample Bayesian inference on the parameters of the proportional hazard models. *Statistics in Medicine* ***16****, 2573-2585.*

[25] Ibrahim, J. G., Chen, M. H., and MacEachern, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canadian Journal of Statistics* ***27****, 701-717.*

[26] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estiamtion and model selection. *International Joint Conference on Artificial Intelligence (IJCAI).* Morgan Kaufmann, 1137-1147.

[27] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* ***22****, 78-86.*

[28] Leblanc, M. and Crowley, J. (1995). Semiparametric regression functions. *Journal of the American Statistical Association* ***90****, 95-105.*

[29] Li, G. (1995). On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statist. & Prob. Letters* ***25****, 95-104.*

[30] Li, G. and Wang, Q. H. (2003). Empirical likelihood regression analysis for right censored data. *Statistica Sinica* ***13****, 51-68.*

[31] Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society* ***B 30****, 31-66.*

[32] Lu, J. C., Chen, D., and Gan, N. C. (2002). Semi-parametric modelling and likelihood estimation with estimating equations. *Australian & New Zealand Journal of Statistics* ***44****, 193-212.*

[33] Mallows, C. L. (1973). Some comments on C_p. *Technometrics* ***15****, 661-675.*

[34] Mallows, C. L. (1995). More comments on C_p. *Technometrics* ***37****, 362-372.*

[35] Monti, A. C. (1997). Empirical likelihood confidence regions in time series models. *Biometrika* ***84****, 395-405.*

[36] Murphy, S. A. (1995). Likelihood ratio-based confidence intervals in survival analysis. *Journal of the American Statistical Association* ***90****, 1399-1405.*

[37] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A* ***135****, 370-384.*

[38] Owen, A. B. (1988). Empirical likelihood ratio confidence interval for a single functional. *Biometrika* ***75****, 237-249.*

[39] Owen, A. B. (1990). Empirical likelihood confidence regions. *Ann. Statist.* ***18****, 90-120.*

[40] Owen, A. B. (1991). Empirical likelihood for linear models. *Ann. Statist.* ***19****, 1725-1747.*

[41] Owen, A. B. (2001). Empirical Likelihood. Chapman and Hall/CRC, New York.

[42] Qin, G. S. and Jing, B. Y. (2001a). Empirical likelihood for censored linear regression. *Scand. J. Statist.* ***28****, 661-673.*

[43] Qin, G. S. and Jing, B. Y. (2001b). Censored partial linear models and empirical likelihood. *Journal of Multivariate Analysis* ***78****, 37-61.*

[44] Qin, G. S. and Tsao, M. (2003). Empirical likelihood inference for median regression models for censored survival data. *Journal of Multivariate Analysis* ***85****, 416-430.*

[45] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics* ***22****, 300-325.*

[46] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* ***6****, 461-464.*

[47] Shen, X. T., Shi, J., and Wong, W. H. (1999). Random sieve likelihood and general regression models. *Journal of the American Statistical Association* ***94****, 835-846.*

[48] Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. *Journal of the Royal Statistical Society B* ***41****, 276-278.*

[49] Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association* ***70****, 865-871.*

[50] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* ***B 58****, 267-288.*

[51] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* ***16****, 385-395.*

[52] Variyath, A. M. (2006). Variable selection in generalized linear models by empirical likelihood. Ph.D. thesis, University of Waterloo, Canada.

[53] Variyath, A. M., Chen, J., and Abraham, B. (2010). Empirical likelihood based variable selection. *Journal of Statistical Planning and Inference* ***140****, 971-981.*

[54] Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* ***56****, 256-262.*

[55] Wang, Q. H. and Li, G. (2002). Empirical likelihood semiparametric regression analysis under random censorship. *Journal of Multivariate Analysis* ***83****, 469-486.*

[56] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* ***9****, 60-62.*