The Pennsylvania State University
The Graduate School

REGULARIZATION PARAMETER SELECTION FOR VARIABLE SELECTION IN HIGH-DIMENSIONAL MODELLING

A Dissertation in
Statistics
by
Yiyun Zhang

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2009

The dissertation of Yiyun Zhang was reviewed and approved* by the following:

Runze Li
Professor of Statistics and Graduate Program Chair
Dissertation Advisor
Chair of Committee

Bing Li
Professor of Statistics

David R. Hunter
Associate Professor of Statistics

Vernon M. Chinchilli
Distinguished Professor and Chair of Health Evaluation Sciences
Professor of Statistics

Bruce G. Lindsay
Willaman Professor of Statistics and Department Head

*Signatures are on file in the Graduate School.

Abstract

Variable selection is an important issue in statistical modelling. Classical approaches select models by applying a penalty related to the size of the candidate model. Exhaustive search is required for these classical methods which is impractical in high-dimensional modelling. Adopting continuous penalties such as the LASSO and the SCAD made it possible to cope with the high-dimensionality. Alike in classical methods, the size of regularization plays a crucial rule in their asymptotic properties. For classical methods, it is well known that AIC-like criteria are asymptotically loss efficient in the sense that they choose the minimum loss model when the true model is infinite dimensional. On the contrary, when there is a finite dimensional correct model, BIC-like criteria are consistent in the sense that they choose the smallest correct model with probability tending to one. Parallel properties for the penalized estimators are studied in this thesis. Extending the results of Wang, Li, and Tsai (2007a), we show that the consistent tuning parameter selector results in a penalized estimator that is also consistent in a general likelihood setting. On the other hand, it is shown that the tuning parameter selector constructed from an efficient criterion is also asymptotically loss efficient for

linear regression. Under the conditions imposed in this thesis, the efficiency result can also be extended to generalized linear models in terms of Kullback-Leibler loss. Our simulation studies suggest the finite sample performances are in line with the theories we present. A real data application is discussed to advocate the use of penalized likelihood variable selection procedures.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I am grateful to my advisor, Dr. Runze Li, for his many helpful ideas and discussion on the contents of this thesis. Second, I want to say thanks to my thesis committee, Dr. Bing Li, Dr. David R. Hunter and Dr. Vernon M. Chinchilli, for their precious time and valuable suggestions in improving the contents of this thesis. I also appreciate many kindnesses from some faculties in my department who were bothered by me during the time I was working on the thesis.

I also want to thank my dear parents and my beloved wife, Yijia Feng. Without their love and support, I cannot finish this paper by myself.

This thesis research has been supported by a National Science Foundation grants DMS 0348869 and National Institute on Drug Abuse grants R21 DA024260 and P50 DA10075. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIDA or the NIH.

Chapter 1

Introduction

1.1 A Brief Overview of the Literature

Model selection has always been an important issue in statistical modelling. It is well known that an overfitted model leads to inefficient estimates of parameters while an underfitted model can result in seriously biased estimates. In practice, a sparse and reasonable model is preferred because of its less complexity and better interpretation. Data analysts often try to find the smallest while adequate collection of data for their model. It is a universal problem in statistics that analysis becomes exponentially more difficult as the dimension of the data increases, i.e. the so called curse of dimensionality. Advancing technology in the scientific world such as genetics, astronomy and computer science allows us to obtain very high dimensional data. This makes the model selection problem more important than it has ever been.

From a technical point of view, the balance between bias and variance is closely related to the variable selection problem. An ideal model should be both accurate (low bias) and precise (low variance). However, if we look at how we construct a

model via least squares or the likelihood principle, our estimator is always tending to include as many predictors as possible which make it less precise in the sense that the variance is enlarged. Then it is clear that, in principle, variable selection can be done with some penalized methods which penalize large models and select a smaller model to achieve the balance of bias and variance.

The earliest attempt on variable selection might be the adjusted-R^2. It was known that the R^2 always increases when a new variable is introduced because of its essence of reducing bias. The adjusted-R^2 was then introduced to correct this. In the 1970's, many variable selection methods were proposed, for example, Mallow's C_p (Mallows, 1973), AIC (Akaike, 1974), BIC (Schwarz, 1978) and so on. These procedures apply to the objective functions some discontinuous penalties (later called entropy penalties) which are proportional to the size of a candidate model. As a consequence, to choose the best model depends on the best-subset-selection, i.e. comparing all possible models according to some criteria and finding the optimal subset. This works relatively well with small dimensional data. As the dimension of the data increases, it becomes impossible to search all possible models exhaustively. Some modifications such as forward/backward selection or stepwise selection are practically useful sometimes but the search is not exhaustive and hence the results may not be optimal. Furthermore, the theoretical properties of these best subset selection models are hard to understand. Reviews of the literature can be found in, e.g., Breiman (1996), McQuarrie and Tsai (1998) and Miller (2002).

In the past two decades, methods with continuous penalties, such as bridge regression (Frank & Friedman, 1993), the LASSO (Tibshirani, 1996) and the SCAD (Fan & Li, 2001), were proposed for variable selection and became popular. Because of the intrinsic advantage of continuity, asymptotic analysis became avail-

able for these continuous penalized estimators. Furthermore, unlike the exhaustive search in classical procedures, fast algorithms are available for the continuous penalty methods. In other words, estimation and variable selection are performed simultaneously via optimizing the penalized objective function.

A natural question follows given all these different variable selection methods. Which of them are optimal? Of course, there is no simple answer to this general question. More specifically, it depends on how "optimality" is defined. In the least squares estimation problem under quadratic loss, there are two asymptotic criteria for comparing variable selection methods in the literature. One criterion focuses on the so called "asymptotic loss efficiency", i.e. how the loss of the suggested estimator compares with the minimum possible loss. The other criterion focuses on the so called "consistency", i.e. how likely the smallest correct model can be obtained. It turned out that these two criteria suggest different strategies for variable selection. Generally speaking, it is shown that some methods, with AIC as an example, are asymptotically loss efficient when the true model is believed to be infinite dimensional (Shibata, 1981; Li, 1987). When there exists a correct model with fixed finite dimension, methods such as BIC are consistent in the sense that they select the best model with high probability. In general, neither AIC-like methods are consistent nor BIC-like methods are efficient (Shao, 1997). There is also little agreement on which criterion is better. Interpretation of these results really depends on the statistical viewpoint of the researcher and the settings in which these methods are applied. A more detailed review follows in the next chapter.

1.2 The Contributions of This Thesis

While all these results are applicable to classical variable selection procedures where entropy penalized methods are used, this dissertation aims to study the continuously penalized methods. Unlike classical AIC or BIC where the size of penalty is fixed, the size of penalty in nonconcave penalized likelihood is determined by the regularization parameter or tuning parameter. Because the property of the resulting estimate is directly determined by this parameter, the choice of tuning parameter is crucial in the nonconcave penalized methodology. Motivated by the existing methods, we proposed a generalized information criterion (GIC) for tuning parameter selection. This criterion covers a wide range of selectors, including AIC and BIC selectors as special cases.

To study the properties of GIC tuning parameter selector, we note that Wang, Li, and Tsai (2007b) showed that SCAD penalized least squares estimate with BIC tuning parameter selector is consistent in variable selection, if we assume the truth to be contained in the candidate model set. However, their results are limited to linear regression with SCAD penalty. In a more general likelihood setting and under some mild conditions, we showed that a wide range of tuning parameter selectors (we call them BIC-type selectors) are able to find the correct true model with probability tending to one. This result completes the discovery by Wang et al. (2007b).

Our second goal is to study if the nonconcave penalized estimates also possess asymptotic loss efficiency when the tuning parameter is properly selected. We show that in the case of linear regression, the AIC tuning parameter selector results in an asymptotically loss efficient penalized estimate under some mild conditions. However, this property does not hold in general if other GIC tuning parameter

selectors are used. This phenomenon is very similar to the classical case where AIC is loss efficient while BIC is not. We further studied the asymptotic loss efficiency of GIC selectors for generalized linear models (GLIM). We show that parallel results still hold for GLIM under the conditions imposed in chapter 4. More specifically, these conditions mainly facilitate the Taylor expansion of MLE for each candidate model.

1.3 The Organization of This Thesis

The dissertation is organized as follows. Existing variable selection methods and their properties are reviewed in chapter 2. A generalized information criterion is proposed in chapter 3 for the selection of tuning parameter. Under some mild conditions, the consistency of penalized likelihood estimators is also studied. Chapter 4 studies how to choose the tuning parameter to achieve asymptotic loss efficiency. Simulation studies are conducted in chapter 5 to assess the performance of the proposed tuning parameter selectors in various settings. A real data example is also analyzed in chapter 5 to advocate the use of nonconcave penalized likelihood methodology. Finally, conclusion remarks and discussions are given in chapter 6.

Chapter 2

Literature Review

The least squares principle and the likelihood principle are the two oldest while most fundamental philosophies in statistics. They coincide in the case of normal regression. Both methods find the estimate via minimizing an objective function (the squared loss or the negative log-likelihood). Regularization of the problem is often achieved by minimizing a penalized objective function. We first review the literature of the variable selection problem in least squares estimation. If we replace the residual sum of squares (RSS) by the negative log-likelihood function, the methodology naturally extends to likelihood settings.

In modern regression modelling, the number of predictors is usually large. This is due to the massive development of technologies and methodologies in bioinformatics, computer science and finance, etc. Another insight comes from nonparametric regression, where the true infinite-dimensional unknown function is approximated by a finite dimensional model. See Fan and Li (2006) for more discussions on the challenges of high dimensional modelling.

These observations suggest that we should allow the number of predictors d to go to infinity as the sample size n increases, instead of fixing the dimension

of parameters. This idea was first systematically introduced in the seminal work of Huber (1973) and then studied by many others. In the literature, a reasonable constraint for d is that $d/n \to 0$, i.e. the number of parameters is small compared to the sample size. Generally speaking, this is a necessary assumption for asymptotic normality (Huber, 1973). In order to establish better theoretical properties, some stronger technical conditions are usually assumed. For example, the oracle property of PLSE requires $d^5/n \to 0$ (Fan & Peng, 2004). By refining the log-likelihood function, this condition can be weakened to $d^3/n \to 0$, which is a reasonable assumption in line with Huber (1973). In the extreme high dimensional case when $d >> n$, the classical statistical procedures cannot apply. Further assumptions and regularization is essential. In this dissertation, we constrain the discussion to the more classical case where $d/n \to 0$. Stronger conditions might apply whenever necessary.

2.1 Regression and Variable Selection in Least Squares Estimation

Consider a regression model

$$E(y_i|\mathbf{x}_i) = \mu(\mathbf{x}_i), \quad i = 1, ..., n, \tag{2.1}$$

where y_i is a scalar response and $\mathbf{x}_i$ is a d-dimensional covariate. The difference between y_i and its mean function is the random error $\epsilon_i = y_i - \mu(\mathbf{x}_i)$.

In a classical (homoscedastic) least squares estimation problem, we further assume ϵ_i's are independent with constant variance σ^2. Our goal is to estimate the mean response $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ with the predictors. In addition, we also want to

estimate the dispersion parameter σ^2 if it is unknown.

Write $\mathbf{y} = (y_1, ..., y_n)^T$ and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$. The mean function is usually modelled by

$$E\mathbf{y} = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{2.2}$$

for some d-dimensional parameter $\boldsymbol{\beta}$. In other words, we assume that $\boldsymbol{\mu}$ lies in the linear space spanned by $\mathbf{X}$. A natural estimate of $\boldsymbol{\mu}$ is the least squares estimate (LSE) minimizing the residual sum of squares

$$RSS = \| \mathbf{y} - \hat{\boldsymbol{\mu}} \|^2 = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})^2. \tag{2.3}$$

Obviously, when more variables are introduced into the model, the RSS can always be improved. However, such improvement is at the expense of higher model complexity. In a variable selection problem, we are interested in whether there exists a smaller set of predictors which also fit the model relatively well without losing much important information. Following the notations in Shao (1997), we define a model as follows.

Definition 2.1 (Candidate Model). *We define α, a subset of*

$$\bar{\alpha} = \{1, ..., d\}, \tag{2.4}$$

as a candidate model, meaning that the corresponding predictors labelled by α are included in the model. Accordingly, $\bar{\alpha}$ is the full model. In addition, we denote the size of model α (i.e., the number of nonzero parameters in α), the corresponding predictor matrix and the coefficients associated with the predictors in model α by d_α, $\mathbf{X}_\alpha$ and $\boldsymbol{\beta}_\alpha$, respectively. Moreover, we denote the collection of all candidate models by $\mathcal{A}$.

Definition 2.2 (Underfitted and Overfitted Models). *If there is a unique true model α_0 in $\mathcal{A}$, that is*

$$\boldsymbol{\mu} = \mathbf{X}_\alpha \boldsymbol{\beta}_\alpha, \tag{2.5}$$

holds for some coefficient $\boldsymbol{\beta}_\alpha$ if and only if $\alpha_0 \subset \alpha$. Therefore, any candidate model $\alpha \not\supset \alpha_0$, is referred to as an underfitted model, while any $\alpha \supset \alpha_0$ other than α_0 itself is referred to as an overfitted model. Denote the set of underfitted and overfitted models by $\mathcal{A}_-$ and $\mathcal{A}_+$ respectively. Furthermore, we say a model is correct if it is either the true model or an overfitted model.

Remark 2.1. It is possible that no model is correct, which means all candidate models are only approximations of the unknown truth. For example, in knots selection for spline regression, the true mean function is in a form of infinite expansion and hence no model is correct.

Remark 2.2. We assume that for fixed n, the size of $\mathcal{A}$ is finite although we allow it grows to infinity as $n \to \infty$. Therefore, if the true model is of infinite dimension, it cannot be included in any $\mathcal{A}$.

Remark 2.3. If there exists more than one correct model with finite dimension, it is reasonable to assume that their intersection is also a correct model. We exclude the cases, for example, when the design matrix is linearly dependent. Therefore we assume that the true model α_0 is unique.

2.2 Variable Selection via Nonconcave Penalized Least Squares

Least squares estimation is simple and possesses nice properties. It is well known that the LSE is the best linear unbiased estimate (BLUE) in a homoscedastic

problem. However, to build a more reliable model in practice, it helps to regularize the problem via penalized least squares estimation (PLSE), i.e. finding $\hat{\boldsymbol{\beta}}$ which minimizes

$$Q(\boldsymbol{\beta}) = \frac{1}{n} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \sum_{j=1}^{d} p_{j,\lambda}(|\beta_j|), \tag{2.6}$$

where $p_{j,\lambda}, j = 1, ..., d_n$ are penalty functions and λ is a tuning parameter which determines the size of regularization.

There are two reasons for regularization. First, the PLSE can be constructed to find a reasonable answer to ill posed problems. Secondly, with carefully chosen penalties, the PLSE are sparse in the sense that many coefficient estimates are exactly zero. The second advantage allows the PLSE to automatically select important variables.

2.2.1 Classical Variable Selection Criteria

There is a vast literature of variable selection, see for example, Akaike (1974), Schwarz (1978), Breiman (1996), Shao (1997), McQuarrie and Tsai (1998), and Miller (2002), among others. For the sake of simplicity, we first study the case with known variance σ^2. When σ^2 is unknown, it usually can be consistently estimated via the full model.

Classical methods are performed via the best subset selection or stepwise regression. Under every candidate model α, the least squares estimate $\hat{\boldsymbol{\mu}}_\alpha = \mathbf{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha$ is first computed. Then the candidate models are compared according to the penalized least squares (PLS) criterion (2.6) where the penalty functions are the so called entropy or L_0 penalties

$$p_\lambda(|\theta|) = \frac{1}{n}\lambda I(\theta \neq 0), \tag{2.7}$$

where $I(\cdot)$ is the indicator function. The PLS criterion can be written as

$$\frac{1}{n} \parallel \mathbf{y} - \hat{\boldsymbol{\mu}}_\alpha \parallel^2 + \frac{1}{n}\lambda_n d_\alpha, \tag{2.8}$$

where d_α is the number of predictors in model α.

Mallow's C_p (Mallows, 1973) is one of the most common criteria in variable selection. It was first introduced by Mallow as an estimate of the risk. It corresponds to the penalty with $\lambda_n = 2\sigma^2$ and selects the model $\hat{\alpha}$ which minimizes

$$C_p(\alpha) = \frac{1}{n} \parallel \mathbf{y} - \hat{\boldsymbol{\mu}}_\alpha \parallel^2 + \frac{2\sigma^2 d_\alpha}{n}. \tag{2.9}$$

Another class of penalties, GIC (Nishii, 1984), corresponds to the entropy penalized criteria with $\lambda_n = \kappa_n \sigma^2$. It selects the model which minimizes

$$GIC(\alpha) = \frac{1}{n} \parallel \mathbf{y} - \hat{\boldsymbol{\mu}}_\alpha \parallel^2 + \frac{\kappa_n \sigma^2 d_\alpha}{n}. \tag{2.10}$$

When $\kappa_n = \log n$, GIC is asymptotically equivalent to the famous Schwarz's Bayesian information criterion (BIC, Schwarz, 1978). On the other hand, when $\kappa_n = 2$, GIC becomes the Akaike's information criterion (AIC, Akaike, 1974), which is equivalent to Mallows' C_p.

There are many other variable selection criteria in the literature corresponding to entropy penalized criteria. These include ϕ-criterion (Hann & Quinn, 1979; Shibata, 1984), RIC (Foster & George, 1994) and small-sample corrected criteria such as AICc (Hurvich & Tsai, 1989).

Note that one serious drawback of these entropy methods is that these procedures often require exhaustive search over every candidate model unless a special algorithm is available. This is computationally too expensive and not practical for

high-dimensional data.

2.2.2 Regularization via Continuous Penalties

In the past decade, many continuous penalties such as the LASSO or L_1 penalty (Tibshirani, 1996), and the SCAD penalty (Fan & Li, 2001) have been proposed for model selection. This allows us to develop some fast algorithms for variable selection instead of the expensive search over all possible models. Moreover, the application of these continuous penalties made it possible to study the asymptotic properties of the PLSEs. See (Fan & Li, 2001) for a review.

The cases with column-orthogonal design matrix (i.e. $\mathbf{X}^T\mathbf{X} = n\mathbf{I}$) have been studied in detail in the literature. In such cases, the forms of the penalized estimates are usually available which are very helpful to understand their performance.

2.2.3 Ridge Regression

The first continuous regularization does not perform variable selection but it lays the foundation for future developments. Ridge regression or the L_2 regularization (Hoerl & Kennard, 1970) minimizes the residual sum of squares (RSS) subject to the constraint $\sum_{j=1}^{d} \beta_j^2 < \tau$. This corresponds to a penalized criterion with a penalty based on L_2 norm (Frank & Friedman, 1993; Fu, 1998). This suggests that we should minimize

$$\frac{1}{n} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \sum_{j=1}^{d} |\beta_j|^2. \tag{2.11}$$

The solution to the above criterion is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.12}$$

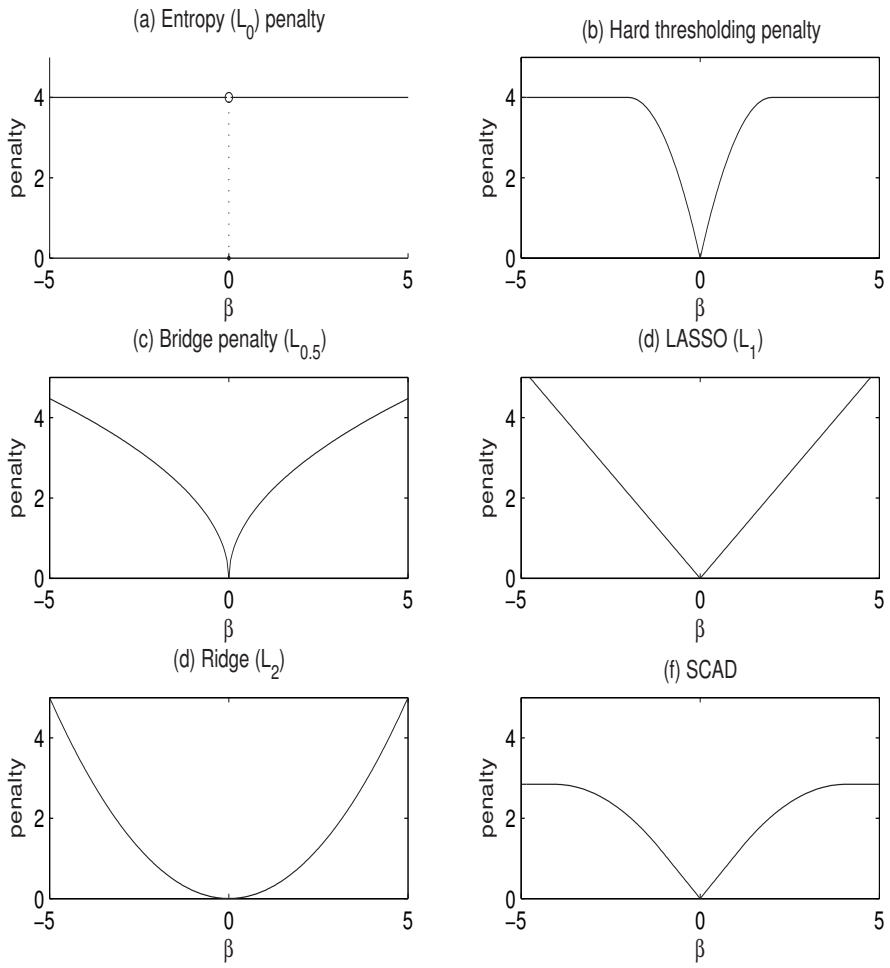

Figure 2.1. Some penalty functions: Entropy, Hard, L_q (q=0.5, 1 and 2) and SCAD.

where $k \geq 0$ is a constant. The implementation of ridge regression is very easy due to this explicit form of its solution.

Ridge regression gives a reasonable solution even if the model is ill posed. From the modelling point of view, it is a shrinkage method which provides regularization and stabilization of the estimates and achieves a favorable bias-variance tradeoff (Breiman, 1996; Fu, 1998).

2.2.4 The Least Absolute Shrinkage and Selection Operator

The L_1 regularization is the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996). It minimizes the RSS subject to the constraint that $\sum_{j=1}^{d} |\beta_j| < \tau$. Therefore it corresponds to a penalized criterion with L_1 penalty and minimizes

$$\frac{1}{n} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \sum_{j=1}^{d} |\beta_j|. \tag{2.13}$$

When $\mathbf{X}$ is orthonormal, the solution to this criterion coincides with the soft thresholding (Donoho & Johnstone, 1994; Donoho, Johnstone, Kerkyacharian, & Picard, 1995)

$$\hat{\beta}_j = \text{sgn}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+, j = 1, ..., d, \tag{2.14}$$

where $\gamma > 0$ is a constant related to the size of the penalty, $\hat{\boldsymbol{\beta}}^0$ is the ordinary least squares estimate and $(\cdot)_+$ refers to the positive part.

It is interesting to compare LASSO with ridge regression and best subset selection. It uses a penalty (L_1) between ridge regression (L_2) and best subset selection (L_0), and its performance also lies between these two in the sense that it is both a selection rule and a shrinkage estimation method. This is a major advantage of LASSO that estimation and model selection are accomplished simultaneously.

It is necessary to mention that penalties between L_0 and L_2 but other than L_1 are also available. General L_q penalties are known as "bridge" regression (Frank

& Friedman, 1993; Tibshirani, 1996; Fu, 1998; Fan & Li, 2001). It minimizes

$$\frac{1}{n} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \sum_{j=1}^{d} |\beta_j|^q. \tag{2.15}$$

The L_q penalties are reviewed in Fu (1998) and Li, Dziak, and Ma (2006).

To implement the LASSO is not as easy as ridge regression because the penalized criterion cannot be optimized explicitly. However the continuity of the penalty function allows us to construct efficient algorithms to optimize the penalized criterion. It can be minimized by linear programming (Tibshirani, 1996), by modified Newton-Raphson methods (Fu, 1998; Fan & Li, 2001) or by the LARS algorithm (Efron, Hastie, Johnstone, & Tibshirani, 2004).

The LASSO is very popular in the literature because of its nice properties and practical implementation. However, the LASSO is always biased due to the shrinkage. Recently Zou (2006) and Leng, Lin, and Wahba (2006) showed that in general the LASSO is not a consistent variable selector in the sense that the proability of overfitting does not approach zero as $n \to \infty$. Fan and Li (2001) gave some intuitive explanation to this issue and proposed another type of penalty that is symmetric and nonconcave on $(0, +\infty)$. These nonconcave penalties, such as SCAD, obtain some nice properties that previous penalties do not have.

2.2.5 The Smoothly Clipped Absolute Deviation

The penalized estimates enjoy better properties with the choice of better penalties. What properties are desirable and what are the conditions for those properties to hold? Fan and Li (2001) addressed this issue and pointed out that a good penalized estimates should achieve asymptotically unbiasedness, sparsity and continuity.

- **Unbiasedness** The penalized estimates should be unbiased when the true parameters are large enough.
- **Sparsity** The penalized estimates should be a variable selection rule that sets small parameters automatically to zero.
- **Continuity** Small perturbation of data should not change the penalized estimates dramatically.

Conditions for these three properties under orthogonal least squares were studied in Fan and Li (2001). It was pointed out that in order for a penalized criterion to satisfy all these three properties, the penalty function has to be nonconcave on $(0, +\infty)$. The Smoothly Clipped Absolute Deviation (SCAD) proposed by Fan and Li (2001) satisfies all these conditions. The SCAD penalty with parameter a and λ has the following as its first derivative:

$$p'_{\lambda,a}(\theta) = \lambda\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\}, \tag{2.16}$$

where $\lambda > 0$ and $a > 2$ are the scale and shape parameters respectively. Fan and Li (2001) pointed out that the penalized estimate is not very sensitive to the choice of a using a Bayesian argument and suggested fixing $a = 3.7$ in applications. Then the SCAD penalty can be written as $p_\lambda(\cdot)$. When $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, the solution to this penalized criterion is given by

$$\hat{\beta}_j = \begin{cases} 0 & \text{if} \quad |\hat{\beta}_j^0| \leq \lambda \\ \hat{\beta}_j^0 - \lambda \operatorname{sgn}(\hat{\beta}_j^0) & \text{if} \quad \lambda < |\hat{\beta}_j^0| \leq 2\lambda \\ \frac{1}{a-2}((a-1)\hat{\beta}_j^0 - a\lambda \operatorname{sgn}(\hat{\beta}_j^0)) & \text{if} \quad 2\lambda < |\beta_j^0| \leq a\lambda \\ \hat{\beta}_j^0 & \text{if} \quad |\hat{\beta}_j^0| > a\lambda \end{cases} \tag{2.17}$$

Fan and Li (2001) also showed that SCAD enjoys an oracle property in that it has sparsity and the parameters are also estimated as if the true model is known in an asymptotic sense.

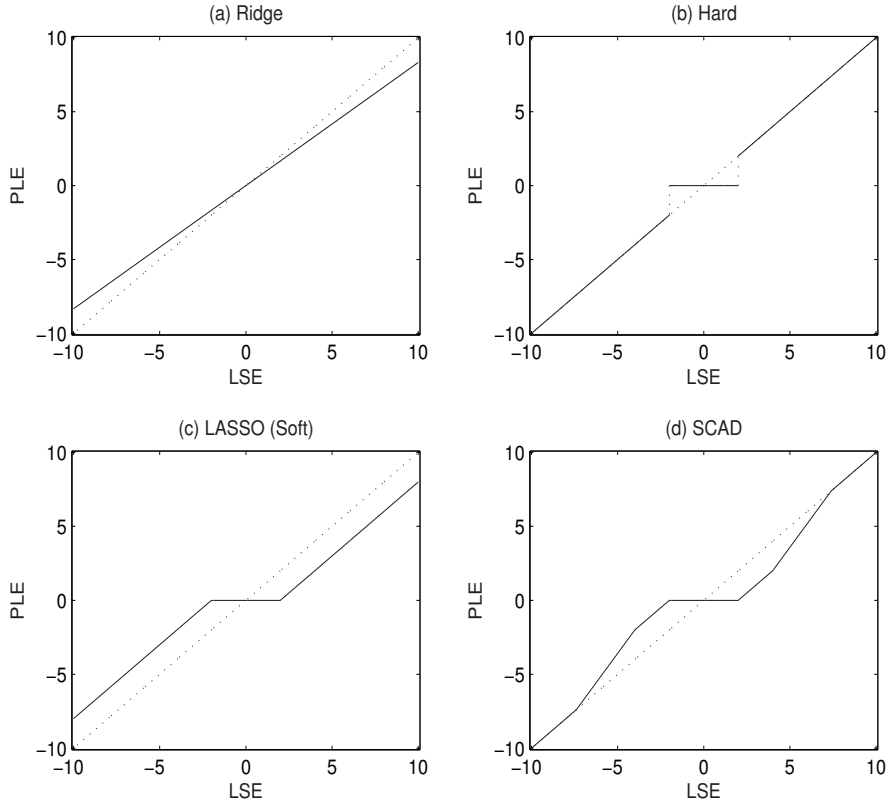

Figure 2.2. Some penalized estimates: column orthogonal linear regression case.

Figure 2.2 gives some insights of the penalized estimates under column orthogonal linear regression. The x-axis corresponds to the ordinary least squares estimate and y-axis corresponds to the penalized least squares estimate. The difference among them are obvious. The ridge estimate shrinks the LSE but is not a thresholding rule and not appropriate for variable selection. The best subset selection (L_0 penalties) or hard-thresholding estimate (Antoniadis, 1997) are unbiased and sparse but discontinuous. The LASSO gives continuous and sparse models but introduces bias. The SCAD gives continuous, sparse and unbiased models.

Many materials are available in the literature for more detailed discussion (Frank & Friedman, 1993; Tibshirani, 1996; Fan & Li, 2001; Li et al., 2006).

2.2.6 The Choice of Tuning Parameter and Degree of Freedom

Despite the appealing theoretical properties of PLSEs with continuous penalties, choosing an appropriate tuning parameter λ remains an important issue both theoretically and practically.

Recall that in classical variable selection procedures, we first calculate the LSE of the candidate models and then compare models with a criterion $Crit(\alpha)$. The choice of tuning parameter also works in this fashion. We can regard the PLS variable selection method as a two-stage penalized method. At the first stage we penalize the original objective function to get the PLSE $\hat{\boldsymbol{\mu}}_\lambda$, and at the second stage we penalize $\| \mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda \|^2$ to choose the tuning parameter.

Step 1: Given a tuning parameter λ, we can find the corresponding PLSE(λ).

Step 2: Compare PLSE(λ) using a criterion $Crit(\lambda)$ constructed from a classical methods, such as AIC/C_p, BIC, GCV, etc.

Step 3: Choose a best tuning parameter which optimizes $Crit(\lambda)$.

For example, it was proposed to choose λ by GCV or BIC criteria (Fan & Li, 2001; Li et al., 2006; Wang et al., 2007b). More specifically, using a GCV selector, we choose $\hat{\lambda}_{GCV}$ such that

$$GCV(\lambda) = \frac{\| \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda \|^2}{n\,(1 - df_\lambda/n)} \tag{2.18}$$

is minimized; or using a BIC selector, we choose $\hat{\lambda}_{BIC}$ such that

$$BIC(\lambda) = \frac{1}{n} \parallel \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda \parallel^2 + \frac{\hat{\sigma}^2 df_\lambda \log n}{n} \tag{2.19}$$

is minimized. Here df_λ is the degrees of freedom of an estimate.

The degrees of freedom was first introduced for a linear smoother $\hat{\boldsymbol{\mu}} = \mathbf{S}\mathbf{y}$ (Hastie & Tibshirani, 1990). As in chapter 4, we will see that the risk of a linear smoother is

$$\frac{\parallel \boldsymbol{\mu} - \mathbf{S}\boldsymbol{\mu} \parallel^2}{n} + \frac{\sigma^2 tr\mathbf{S}}{n}. \tag{2.20}$$

When $\mathbf{S}$ is the projection $\mathbf{H}_\alpha$ corresponding to a model α, $tr\mathbf{H}_\alpha = d_\alpha$ which is also the number of predictors in this model. This gives a natural definition of degrees of freedom as $df = tr\mathbf{S}$. See Zou, Hastie, and Tibshirani (2007) for more discussion of the degrees of freedom.

However, an estimate might not be a linear smoother in most cases. Then the degrees of freedom can be defined in various ways. One naive way to model the degrees of freedom is using the number of nonzero predictors

$$df_N = \sum_{j=1}^{d} I(\hat{\beta}_j \neq 0). \tag{2.21}$$

Two other degrees of freedoms are used in the literature:

$$df_S = \sum_{j=1}^{d} |\hat{\beta}_j|/|\hat{\beta}_j^{LS}|, \tag{2.22}$$

where $\hat{\beta}_j^{LS}$'s and $\hat{\beta}_j$'s are the LSE and PLSE respectively; and

$$df_L = tr\left(\mathbf{X}_{\hat{\alpha}_\lambda}\left(\mathbf{X}_{\hat{\alpha}_\lambda}^T\mathbf{X}_{\hat{\alpha}_\lambda} + n\Sigma_\lambda\right)^{-1}\mathbf{X}_{\hat{\alpha}_\lambda}^T\right) \tag{2.23}$$

where $\Sigma_\lambda = \text{diag}_{\hat{\beta}_j \neq 0} \left\{ p'_\lambda(|\hat{\beta}_j|)/|\hat{\beta}_j| \right\}$. df_S is constructed from a shrinkage point of view. df_L adopts the idea of degrees of freedom for a linear smoother, although the estimate is not really linear. In linear regression models, Efron et al. (2004) analyzed the expansion of model error and argued that the degrees of freedom should be

$$df_E = \sum_{i=1}^{n} Cov(\hat{\mu}_i, y_i)/\sigma^2. \tag{2.24}$$

Zou et al. (2007) later suggested using d_{α_λ} by showing that d_{α_λ} is an unbiased estimator of the df_E.

2.3 Criteria of Good Variable Selection Methods

Generally speaking, there are two asymptotic criteria for comparing variable selection methods in the literature, efficiency and consistency (Shibata, 1981; Nishii, 1984; Li, 1987; Shao, 1997). Recently there are some studies that focus on the finite sample minimax properties (Yang & Barron, 1999; Barron, Birgé, & Massart, 1999; Birgé & Massart, 2001). We will focus the asymptotic efficiency and the consistency criterion in this dissertation. A brief discussion of the minimax approach will be given in the end of this section.

2.3.1 Efficiency Criterion

Definition 2.3 (Squared Loss and Risk). *The average squared loss of an estimate $\hat{\boldsymbol{\beta}}$ is defined by*

$$L(\hat{\boldsymbol{\beta}}) = \frac{\| \boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}} \|^2}{n}. \tag{2.25}$$

The risk of this estimate is defined as the expected squared loss

$$R(\hat{\boldsymbol{\beta}}) = E\frac{\| \boldsymbol{\mu} - \mathbf{X}\hat{\boldsymbol{\beta}} \|^2}{n}. \tag{2.26}$$

Many researchers consider the true model is of infinite dimension and therefore any finite dimensional model only approximates the truth. Following this philosophy, the model selected should be as close as possible to the truth. In a least squares problem where the mean function is estimated by the LSE, a model selection method is considered to be asymptotically loss efficient if the loss of the model selected, $L_n(\hat{\alpha})$, is the minimum among all $\alpha \in \mathcal{A}_n$.

Definition 2.4 (Asymptotic Loss Efficiency for Classical Criteria). *A classical variable selection criterion is said to be asymptotically loss efficient if the model selected $\hat{\alpha}$ satisfies*

$$\frac{L_n(\hat{\boldsymbol{\beta}}^*_{\hat{\alpha}})}{\inf_{\alpha\in\mathcal{A}_n} L_n(\hat{\boldsymbol{\beta}}^*_{\alpha})} \to 1, \tag{2.27}$$

in probability, where $\hat{\boldsymbol{\beta}}^*_{\alpha}$ is the least squares estimate under model α.

This concept was first introduced by Shibata (1981) and then studied by Li (1987) and Shao (1997). It is shown that FPE, AIC, C_p and GCV are efficient. Many attempts have been made to find small sample correction to these asymptotic criteria, for example, the AICc criterion (Hurvich & Tsai, 1989) is one well-known corrected version of AIC.

As for penalized estimates, the size of the tuning parameter determines the final model. Therefore, we can define the asymptotic loss efficiency of a tuning parameter selection procedure accordingly.

Definition 2.5 (Asymptotic Loss Efficiency for Penalized Estimates). *A*

tuning parameter selection procedure is said to be asymptotically loss efficient if

$$\frac{L(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})}{\inf_{\lambda\in[0,\lambda_{\max}]} L(\hat{\boldsymbol{\beta}}_{\lambda})} \to 1, \tag{2.28}$$

in probability, where $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ is the penalized estimate associated with the tuning parameter $\hat{\lambda}$ selected by this procedure.

2.3.2 Consistency Criterion

Many other researchers assume that there exists a finite dimensional true model in the set of the candidate models. Following such a philosophy, our goal should be recognizing the true model from the candidates. A model selection method is said to be consistent if it selects the true model with high probability. This criterion is related to the concept of sparsity. We want to find a model that is correct and as sparse as possible. BIC (or its generalization GIC) is a representative of consistent methods.

To link a general estimate with the consistency criterion, we define the consistency of an estimate in variable selection as follows.

Definition 2.6 (Consistency in Variable Selection). *If there exists at least one correct model and α_0 is the true model, a model selection criterion is consistent if*

$$P(\hat{\alpha} = \alpha_0) \to 1, \tag{2.29}$$

where $\hat{\alpha}$ is the model selected.

Efficient classical variable selection criteria such as AIC are not consistent. It was shown that there is positive probability for AIC to overfit. Consistent classical variable selection criteria, such as BIC or GIC with $\kappa_n \to \infty$, are also

not efficient in general. In some special cases, GIC is efficient, for example when a finite dimensional true model exists (Shao, 1997). However, this fundamental assumption is questionable in practice. First it is hard to argue the existence of the true model. Even if we assume there is a true model, it might not be of finite dimension.

2.3.3 Minimax-Rate Approach

Both the efficiency criterion and the consistency criterion are asymptotic criteria. Although some small sample corrections such as AICc are available, they are still far from exact. It is possible that the small sample behavior of an estimate might be very different from what is suggested by its asymptotic property. Recently, researchers began to study the minimax risk behavior of variable selection procedures. Unlike the asymptotic efficiency and consistency, it focuses on the finite sample risk. More specifically, it would be ideal if we can find an estimate $\hat{\boldsymbol{\mu}}$ that is minimax risk optimal, i.e. the risk $R(\hat{\mu})$ achieves $\inf_{\hat{\mu}\in\mathcal{F}} \sup_{\mu} R(\hat{\mu})$, the minimum risk among all possible estimates. This turns out to be too ambitious. Donoho et al. (1995) laid a foundation for post-classical minimax analysis. Following their idea, in those cases where it is impossible to find a minimax estimate, we could instead try to find an estimate that is minimax-rate optimal, i.e.

$$\frac{R(\hat{\mu})}{C_n \inf_{\hat{\mu}\in\mathcal{F}} \sup_{\mu} R(\hat{\mu})} \to 1, \quad n \to \infty, \tag{2.30}$$

where C_n is a nonrandom number which only depends on n.

Yang and Barron (1999) showed that AIC is minimax-rate optimal. On the other hand, BIC does not possess minimax optimality. It is worth noticing that although AIC is minimax-rate optimal, it pays a big price for this property that

it overfits the model with probability bounded from zero. In a recent study, Yang (2005) argued that there is no hope to construct an estimate that is both consistent and minimax-rate optimal.

We restrain our focus on the asymptotic loss efficiency and consistency, so the study of minimax-rate optimality of PLSE is beyond the scope of this dissertation.

Chapter 3

Consistency

In many problems, we assume the true model is a smaller model contained in the full candidate model. Under this framework, it is natural to ask whether a variable selection procedure is able to identify the true model. We say a procedure is consistent in variable selection if the truth can be discovered with high probability as sample size increases. Motivated by the generalized information criterion used in linear regression variable selection, we propose a GIC tuning parameter selector to choose the regularization parameter λ for penalized likelihood functions. It is of great interest to study the consistency of GIC tuning parameter selectors.

3.1 Generalized Information Criterion

3.1.1 Extending Classical GIC

In the normal linear regression model, $y_i = \mathbf{x}_i^T \boldsymbol{\beta}_\alpha + e_i$ and $e_i \sim N(0, \sigma^2)$ for $i = 1, \cdots, n$, Nishii (1984) proposed the generalized information criterion for classical

variable selections. It is

$$\mathrm{GIC}_{\kappa_n}(\alpha) = \log \hat{\sigma}^2_\alpha + \frac{1}{n}\kappa_n d_\alpha, \tag{3.1}$$

where d_α is the number of parameters in model α, $\boldsymbol{\beta}_\alpha$ is the parameter of the candidate model α, $\hat{\sigma}^2_\alpha$ is the maximum likelihood estimator of σ^2, and κ_n is a positive number that controls the properties of variable selection. Note that Nishii's GIC is different from the GIC proposed by Konishi and Kitagawa (1996). When $\kappa_n = 2$, GIC becomes AIC, while $\kappa_n = \log n$ leads to GIC being BIC. Because GIC contains a broad range of selection criteria, this motivates us to propose the following GIC regularization parameter selector,

$$\mathrm{GIC}_{\kappa_n}(\lambda) = \frac{1}{n}\{G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda) + \kappa_n df_\lambda\}, \tag{3.2}$$

where $G(\mathbf{y}, \hat{\boldsymbol{\beta}}_\lambda)$ measures the fitting of model α_λ, $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\hat{\boldsymbol{\beta}}_\lambda$ is the penalized parameter estimator obtained by maximizing Equation (3.11) with respect to $\boldsymbol{\beta}$, and df_λ is the degrees of freedom of model α_λ. For any given κ_n, we select λ that minimizes $\mathrm{GIC}_{\kappa_n}(\lambda)$.

Remark 3.1. For any given model α, we are able to obtain the non-penalized parameter estimator $\hat{\beta}^*_\alpha$ by maximizing the log-likelihood function $\ell(\boldsymbol{\beta})$ in (3.11). Then, Equation (3.2) becomes

$$\mathrm{GIC}^*_{\kappa_n}(\alpha) = \frac{1}{n}\{G(\mathbf{y}, \hat{\boldsymbol{\beta}}^*_\alpha) + \kappa_n d_\alpha\}, \tag{3.3}$$

which can be used for classical variable selections. In addition, $\mathrm{GIC}^*_{\kappa_n}(\alpha)$ turns into Nishii's $\mathrm{GIC}_{\kappa_n}(\alpha)$ if we replace $G(\mathbf{y}, \hat{\boldsymbol{\beta}}^*_\alpha)$ in (3.3) with the $-2\log$-likelihood function of the fitted normal regression model.

3.1.2 Technical Conditions

To investigate the theoretical properties of the GIC selector, we introduce the technical conditions given below.

(C1) The upper limit of the regularization parameter, $\lambda_{\max}$, depends on n and satisfies $\lambda_{\max} \to 0$ as $n \to \infty$.

(C2) There exists a constant m such that the penalty $p_\lambda(\theta)$ satisfies $p'_\lambda(\theta) = 0$ for $\theta > m\lambda$.

(C3) If $\lambda_n \to 0$ as $n \to \infty$, then the penalty function satisfies

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} \sqrt{n} p'_{\lambda_n}(\theta) \to \infty. \tag{3.4}$$

(C4) For any candidate model $\alpha \in \mathcal{A}$, there exists $c_\alpha > 0$ such that $\frac{1}{n} G(\mathbf{y}, \hat{\boldsymbol{\beta}}^*_\alpha) \xrightarrow{P} c_\alpha$. In addition, for any underfitted model $\alpha \in \mathcal{A}$, $c_\alpha > c_{\alpha_0}$, where c_{α_0} is the limit of $\frac{1}{n} G(\mathbf{y}, \boldsymbol{\beta}_{\alpha_0})$ and $\boldsymbol{\beta}_{\alpha_0}$ is the parameter vector of the true model α_0.

Condition (C1) indicates that a smaller regularization parameter is needed if the sample size is large. Condition (C2) assures that the resulting penalized likelihood estimate is asymptotically unbiased (Fan & Li, 2001). Both the SCAD and Zhang's (2007) minimax concave penalties satisfy this condition. Condition (C3) is adapted from Fan and Li's (2001) Equation (3.5), which is used to study the oracle property. Condition (C4) assures that the underfitted model yields a larger measure of model fitting than that of the true model.

Remark 3.2. Note that we can partition the tuning parameter interval $[0, \lambda_{\max}]$ into the underfitted, true, and overfitted subsets, respectively,

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

$$\begin{aligned}\Omega_0 &= \{\lambda : \alpha_\lambda = \alpha_0\}, \text{ and} \\ \Omega_+ &= \{\lambda : \alpha_\lambda \supset \alpha_0 \text{ and } \alpha_\lambda \neq \alpha_0\}.\end{aligned}$$

Here α_λ is the model associated with tuning parameter λ. These partitions allow us to assess the consistency of regularization parameter selectors later.

3.1.3 Degrees of Freedom

We next study the degrees of freedom used in the second term of GIC. In the selection of the regularization parameter, Fan and Li (2001, 2002) proposed that the degrees of freedom be the trace of the approximate linear projection matrix, i.e.

$$df_L(\lambda) \triangleq tr\left\{\left(\nabla^2_\lambda Q^*(\hat{\boldsymbol{\beta}}_\lambda)\right)^{-1}\nabla^2_\lambda \ell(\hat{\boldsymbol{\beta}}_\lambda)\right\}, \tag{3.5}$$

where $Q^*(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^d q_\lambda(|\beta_j|)$, $[\nabla^2_\lambda Q^*(\boldsymbol{\beta})]_{jj'} = \frac{\partial^2}{\partial\beta_j\beta_{j'}}Q^*(\boldsymbol{\beta})$, and $[\nabla^2_\lambda \ell(\boldsymbol{\beta})]_{jj'} = \frac{\partial^2}{\partial\beta_j\beta_{j'}}\ell(\boldsymbol{\beta})$ for j, j' such that $\hat{\beta}_j \neq 0$ and $\hat{\beta}_{j'} \neq 0$. To understand the large sample property of $df_L(\lambda)$, we show its asymptotic behavior given below.

Proposition 3.1. *Assume that the penalized likelihood estimator $\hat{\boldsymbol{\beta}}_\lambda$ is sparse (i.e., with probability tending to one, $\hat{\beta}_{\lambda j} = 0$ if the true value of β_j is 0) and consistent, where $\hat{\beta}_{\lambda j}$ is the j-th component of $\hat{\boldsymbol{\beta}}_\lambda$. Under conditions (C1) and (C2), we have*

$$P\left\{df_L(\lambda) = d_{\alpha_\lambda}\right\} \to 1, \tag{3.6}$$

where d_{α_λ} is the size of model α_λ.

Proof. After algebraic simplifications,

$$df_L(\lambda) = tr\left\{\left(\nabla^2_\lambda \ell_{\alpha_\lambda}(\hat{\boldsymbol{\beta}}_\lambda) + n\Sigma_\lambda\right)^{-1}\nabla^2_\lambda \ell(\hat{\boldsymbol{\beta}}_\lambda)\right\}, \tag{3.7}$$

where $\Sigma_\lambda = \text{diag}_{\hat{\beta}_{\lambda j}\neq 0}\left\{p'_\lambda(|\hat{\beta}_{\lambda j}|)/|\hat{\beta}_{\lambda j}|\right\}$. Because of the consistency and sparsity of $\hat{\boldsymbol{\beta}}_\lambda$, $\hat{\beta}_{\lambda j}$ converges to β_j with probability tending to 1 for all j such that $\hat{\beta}_{\lambda j} > 0$. Hence, those $\hat{\beta}_{\lambda j}$ are all bounded from 0. This result, together with conditions (C1) and (C2), implies that $\Sigma_\lambda = \mathbf{0}$ with probability tending to 1. Subsequently, using the fact that $n^{-1}\nabla^2_\lambda \ell(\hat{\boldsymbol{\beta}}_\lambda) = O_P(1)$, we complete the proof. □

The above proposition suggests that the difference between $df_L(\lambda)$ and the size of the model, d_{α_λ}, is small. Because d_{α_λ} is simple to calculate, we use it as the degrees of freedom df_λ in (3.2). In linear regression models, Efron et al. (2004) and Zou et al. (2007) also suggested using d_{α_λ} as an estimator of the degrees of freedom for LASSO. Moreover, Zou et al. (2007) showed that d_{α_λ} is an asymptotically unbiased estimator. In this article, our asymptotical results are valid without regard to the use of $df_L(\lambda)$ or d_{α_λ} as the degrees of freedom. When the sample size is small, however, $df_L(\lambda)$ should be considered.

3.2 Generalized Linear Models

3.2.1 GIC Tuning Parameter Selector for GLIM

Consider the generalized linear model (GLIM, see McCullagh & Nelder, 1989), whose conditional density function of y_i given x_i is

$$f_i(y_i;\theta_i,\phi) = \exp\left\{\left[y_i\theta_i - b(\theta_i)\right]/a(\phi) + c(y_i,\phi)\right\}, \tag{3.8}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are suitably chosen functions, θ_i is the canonical parameter, $E(y_i|x_i) = \mu_i = b'(\theta_i)$, $g(\mu_i) = \theta_i$, g is a link function, and ϕ is a scale parameter. Throughout this paper, we assume that ϕ is known (such as the logistic

regression model and the Poisson log-linear model) or that it can be estimated by fitting the data with the full model (for instance, the normal linear model). In addition, we follow classical regression theory and model θ_i by $\mathbf{x}_i^T\boldsymbol{\beta}$. Based on (3.8), the log likelihood-based function in (3.11) is

$$\ell(\boldsymbol{\beta}) = \ell(\boldsymbol{\mu};\mathbf{y}) = \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(y_i;\theta_i,\phi) = \sum_{i=1}^{n} [\{y_i\mathbf{x}_i^T\boldsymbol{\beta} - b(\mathbf{x}_i^T\boldsymbol{\beta})\}/a(\phi) + c(y_i,\phi)], \tag{3.9}$$

where $\boldsymbol{\mu} = (\mu_1,\cdots,\mu_n)^T$ and $\mathbf{y} = (y_1,\cdots,y_n)^T$. Then, the resulting scaled deviance of a penalized estimate $\hat{\boldsymbol{\beta}}_\lambda$ is

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}_\lambda) = 2\{\ell(\mathbf{y};\mathbf{y}) - \ell(\hat{\boldsymbol{\mu}}_\lambda;\mathbf{y})\}, \tag{3.10}$$

where $\hat{\boldsymbol{\mu}}_\lambda = (g^{-1}(\mathbf{x}_1^T\hat{\boldsymbol{\beta}}_\lambda),\cdots,g^{-1}(\mathbf{x}_n^T\hat{\boldsymbol{\beta}}_\lambda))^T$.

Adopting Fan and Li's (2001) approach, we define a penalized likelihood to be

$$Q(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{3.11}$$

where $p_\lambda(\cdot)$ and λ are the penalty function and regularization parameter as discussed in chapter 2. With properly chosen tuning parameter λ, the resulting penalized estimate sets small coefficients to zero automatically, and hence variable selection is achieved.

The GIC tuning parameter selector can be applied to select λ. For model α_λ, we employ the scaled deviance $D(\mathbf{y};\hat{\boldsymbol{\mu}}_\lambda)$ as the goodness-of-fit measure, $G(\mathbf{y},\hat{\boldsymbol{\beta}}_\lambda)$, in (3.2) so that the resulting generalized information criterion for GLIM is

$$\mathrm{GIC}_{\kappa_n}(\lambda) = \frac{1}{n}D(\mathbf{y};\hat{\boldsymbol{\mu}}_\lambda) + \frac{1}{n}\kappa_n df_\lambda. \tag{3.12}$$

In addition, when we fit the data with the non-penalized likelihood approach under model α, GIC becomes

$$\mathrm{GIC}^*_{\kappa_n}(\alpha) = \frac{1}{n} D(\mathbf{y}; \hat{\boldsymbol{\mu}}^*_\alpha) + \frac{1}{n} \kappa_n d_\alpha, \tag{3.13}$$

where $\hat{\boldsymbol{\mu}}^*_\alpha = (g^{-1}(\mathbf{x}_1^T \hat{\boldsymbol{\beta}}^*_\alpha), \cdots, g^{-1}(\mathbf{x}_n^T \hat{\boldsymbol{\beta}}^*_\alpha))^T$, and $\hat{\boldsymbol{\beta}}^*_\alpha$ is the non-penalized maximum likelihood estimator of β. Accordingly, GIC* can be used in classical variable selection (see Eq. (3.10) of McCullagh & Nelder, 1989).

3.2.2 Consistency of GIC Tuning Parameter Selectors

Theorem 3.1. *Assume that the technical condition (C4) holds.*

(A) If there exists a positive constant M such that $\kappa_n < M$, then the tuning parameter λ selected by minimizing $GIC_{\kappa_n}(\lambda)$ in (3.12) satisfies

$$P\{\lambda \in \Omega_-\} \to 0, \text{ and } P\{\lambda \in \Omega_+\} \geq \pi, \tag{3.14}$$

where π is a nonzero probability.

(B) Suppose that conditions (C1)-(C3) are satisfied. If $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$, then the tuning parameter λ selected by minimizing $GIC_{\kappa_n}(\lambda)$ in (3.12) satisfies $P\{\alpha_\lambda = \alpha_0\} \to 1$.

Remark 3.3. Theorem 3.1 provides guidance on the choice of the regularization parameter. Because $\kappa_n = 2$ satisfies the boundedness condition in Theorem 3.1(A), we name GIC with the bounded κ_n the AIC-type selector. In contrast, because $\kappa_n = \log n$ fulfills the conditions of Theorem 3.1(B), we call GIC with $\kappa_n \to \infty$ and $\kappa_n/\sqrt{n} \to 0$ the BIC-type selector. Accordingly, Theorem 3.1(A) implies that

the AIC-type selector tends to overfit without regard to which penalty function being used. However, Theorem 3.1(B) indicates that BIC-type selector enables us to identify the true model consistently. Thus, the nonconcave penalized likelihood of the generalized linear model with the BIC-type selector possesses the oracle property.

Remark 3.4. In linear regression models, Wang et al. (2007b) demonstrated that Fan and Li's (2001) GCV-selector for the SCAD penalized least squares procedure cannot select the tuning parameter satisfactorily. They further proposed the following BIC tuning parameter selector,

$$\mathrm{BIC}^*_\lambda = \log(\hat{\sigma}^2_\lambda) + \frac{1}{n}\log(n)df_L(\lambda), \tag{3.15}$$

where $\hat{\sigma}^2_\lambda = n^{-1}\sum_{i=1}^n (y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\lambda)^2$. Using the result of $\log(1+t) \approx t$ for small t, BIC^*_λ is approximately equal to

$$\mathrm{BIC}^{**}_\lambda = \frac{1}{n}D^{**}(\mathbf{y};\hat{\boldsymbol{\mu}}_\lambda) + \frac{1}{n}\log(n)df_L(\lambda), \tag{3.16}$$

where $D^{**}(\mathbf{y};\hat{\boldsymbol{\mu}}_\lambda)\hat{\sigma}^2_\lambda/\hat{\sigma}^2_{\tilde{\alpha}}$ is the scaled deviance of normal distribution, and $\hat{\sigma}^2_{\tilde{\alpha}}$ is the dispersion parameter estimator computed from the full model. It can be seen that $\mathrm{BIC}^{**}_\lambda$ is a BIC-type selector. Under the conditions in Theorem 3.1(B), the SCAD penalized least squares procedure with $\mathrm{BIC}^{**}_\lambda$-selector possesses the oracle property, which is consistent with the findings in Wang et al. (2007b).

Before proving Theorem 3.1, we show the following two lemmas. Then, Theorems 3.1(A) and 3.1(B) follow from Lemmas 3.1 and 3.2, respectively. For the sake of convenience, we denote $D(\mathbf{y};\hat{\boldsymbol{\mu}}_\lambda)$ in (3.12) and $D(\mathbf{y};\hat{\boldsymbol{\mu}}^*_\alpha)$ in (3.13) by $D(\lambda)$ and $D^*(\alpha)$, respectively.

Lemma 3.1. *Assume that there exists a positive constant M such that $\kappa_n < M$. Then under condition (C4), we have*

$$P\left\{\inf_{\lambda\in\Omega_-} GIC_{\kappa_n}(\lambda) > GIC^*_{\kappa_n}(\bar{\alpha})\right\} \to 1 \tag{3.17}$$

and

$$\liminf_{n\to\infty} P\left\{\inf_{\lambda\in\Omega_0} GIC_{\kappa_n}(\lambda) > GIC^*_{\kappa_n}(\bar{\alpha})\right\} \geq \pi. \tag{3.18}$$

as $n \to \infty$.

Proof. For the model α_λ, the non-penalized maximum likelihood estimator, $\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}$, maximizes $\ell(\boldsymbol{\beta})$. This implies $D(\lambda) \geq D^*(\alpha_\lambda)$, which leads to

$$\mathrm{GIC}_{\kappa_n}(\lambda) = D(\lambda)/n + \kappa_n df_\lambda/n > D^*(\alpha_\lambda)/n. \tag{3.19}$$

Then, in conjunction with (3.19), we obtain that

$$\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) > \frac{D^*(\alpha_\lambda)}{n} - \frac{D^*(\bar{\alpha})}{n} - \frac{\kappa_n d_{\bar{\alpha}}}{n}, \tag{3.20}$$

holds true for any $\lambda \in \Omega_- = \{\lambda : \alpha_\lambda \not\supset \alpha_0\}$. Together with condition (C4) and the fact that $\kappa_n < M$, so that $\kappa_n d_{\bar{\alpha}}/n = o_P(1)$,

$$\begin{aligned} P\left\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) > 0\right\} &\geq P\left\{\min_{\alpha\not\supset\alpha_0} \frac{D^*(\alpha)}{n} - \frac{D^*(\bar{\alpha})}{n} - \frac{\kappa_n d_{\bar{\alpha}}}{n} > 0\right\} \\ &= P\left\{\min_{\alpha\not\supset\alpha_0} c_\alpha - c_{\bar{\alpha}} + o_P(1) > 0\right\} \to 1, \end{aligned} \tag{3.21}$$

as $n \to \infty$. The last step used the fact that although $\lambda \in \Omega_- \subset [0, \lambda_{\max}]$, the number of underfitting models is finite, and hence under condition (C4), $\min_{\alpha\not\supset\alpha_0} c_\alpha$ is strictly greater than $c_{\bar{\alpha}}$ which equals c_{α_0} from classical theory. The above equation

yields (3.17) immediately.

Next we show that model selected by the AIC-type selector is overfitted with a positive probability. For any $\lambda \in \Omega_0$, $\alpha_\lambda = \alpha_0$. Then subtracting $\mathrm{GIC}^*_{\kappa_n}(\bar{\alpha})$ from the both sides of (3.19) and subsequently taking $\inf_{\lambda\in\Omega_0}$ over $\mathrm{GIC}_{\kappa_n}(\lambda)$, we have

$$\inf_{\lambda\in\Omega_0} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) > D^*(\alpha_0)/n - \left[D^*(\bar{\alpha})/n + \kappa_n d_{\bar{\alpha}}/n\right]. \tag{3.22}$$

Note that the right hand side of the above equation does not involve λ which will eventually be selected by data driven methods. It is known that under standard regularity conditions, $D^*(\alpha_0) - D^*(\bar{\alpha}) = -2\left[\ell(\hat{\boldsymbol{\beta}}^*_{\alpha_0}) - \ell(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})\right] \xrightarrow{\mathcal{L}} \chi^2_{d_{\bar{\alpha}}-d_{\alpha_0}}$, where $\hat{\boldsymbol{\beta}}^*_{\alpha_0}$ and $\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}}$ are the non-penalized maximum likelihood estimators computed from the true and full models, respectively. This result, together with $\kappa_n < M$ and (3.22), yields

$$\begin{aligned} P\left\{\inf_{\lambda\in\Omega_0} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}^*_{\kappa_n}(\bar{\alpha}) > 0\right\} &\geq P\left\{-\frac{2}{n}\left[\ell(\hat{\boldsymbol{\beta}}^*_{\alpha_0}) - \ell(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})\right] - \kappa_n d_{\bar{\alpha}}/n > 0\right\} \\ &\geq P\left\{-2\left[\ell(\hat{\boldsymbol{\beta}}^*_{\alpha_0}) - \ell(\hat{\boldsymbol{\beta}}^*_{\bar{\alpha}})\right] > d_{\bar{\alpha}} M\right\} \\ &\to P\left\{\chi^2_{d_{\bar{\alpha}}-d_{\alpha_0}} \geq d_{\bar{\alpha}} M\right\} \triangleq \pi. \end{aligned} \tag{3.23}$$

This implies(3.18), and we complete the proof of Lemma 1. □

Lemma 3.2. *Assume conditions (C1)—(C4) hold, and let* $\lambda_n = \kappa_n/\sqrt{n}$. *If* κ_n *satisfies* $\kappa_n \to \infty$ *and* $\lambda_n \to 0$ *as* $n \to \infty$, *then*

$$P\left\{GIC_{\kappa_n}(\lambda_n) = GIC^*_{\kappa_n}(\alpha_0)\right\} \to 1, \tag{3.24}$$

and

$$P\left\{\inf_{\lambda\in\Omega_-\cup\Omega_+} GIC_{\kappa_n}(\lambda) > GIC_{\kappa_n}(\lambda_n)\right\} \to 1. \tag{3.25}$$

Proof. Without loss of generality, we assume that the first d_{α_0} coefficients of $\boldsymbol{\beta}_{\alpha_0}$ in the true model are nonzero and the rest are zeros. Note that the density function of the generalized linear model satisfies Fan and Li's (2001) three regularity conditions (A)—(C). These conditions, together with condition (C3) and the assumptions of κ_n stated in Lemma 3.2, allow us to apply Fan and Li's Theorems 1 and 2 to show that, with probability tending to 1, the last $d - d_{\alpha_0}$ components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ are zeros and the first d_{α_0} components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ satisfy the normal equations

$$\frac{\partial}{\partial \beta_j}\ell(\hat{\boldsymbol{\beta}}_{\lambda_n}) + b_{\lambda_n j} = 0 \; for \; j = 1, ..., d_{\alpha_0}, \tag{3.26}$$

where $b_{\lambda_n j} = p'_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|)\text{sgn}(\hat{\beta}_{\lambda_n j})$, and $\hat{\beta}_{\lambda_n j}$ is the j-th component of $\hat{\boldsymbol{\beta}}_{\lambda_n}$.

Using the oracle property, we have $|\hat{\beta}_{\lambda_n j}| \to |\beta_j| \geq \min_{1 \leq j \leq d_{\alpha_0}} |\beta_j|$. Then, under conditions (C1) and (C2), there exists a constant m so that $p'_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|) = 0$ for $\min_{1 \leq j \leq d_{\alpha_0}} |\beta_j| > m\lambda_n$ as n gets large. Accordingly, $P(b_{\lambda_n j} = 0) \to 1$ for $j = 1, ..., d_{\alpha_0}$. This together with (3.26) implies that, with probability tending to 1, the first d_{α_0} components of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ solve the normal equations

$$\frac{\partial}{\partial \beta_j}\ell(\hat{\boldsymbol{\beta}}_{\lambda_n}) = 0, \;\; j = 1, ..., d_{\alpha_0}, \tag{3.27}$$

and the remaining $d - d_{\alpha_0}$ components are zeros. This is exactly the same as the normal equation in solving the non-penalized maximum likelihood estimator $\hat{\boldsymbol{\beta}}^*_{\alpha_0}$. As a result, $\hat{\boldsymbol{\beta}}^*_{\alpha_0} = \hat{\boldsymbol{\beta}}_{\lambda_n}$ with probability tending to 1. It follows that

$$P\left\{D(\lambda_n) = D^*(\alpha_0)\right\} = P\left\{\ell(\hat{\boldsymbol{\beta}}_{\lambda_n}) = \ell(\hat{\boldsymbol{\beta}}^*_{\alpha_0})\right\} \to 1. \tag{3.28}$$

Moreover, using the result from Proposition 3.1, we have $P\left\{df_{\lambda_n} = d_{\alpha_0}\right\} \to 1$.

Consequently,

$$P\left\{\mathrm{GIC}_{\kappa_n}(\lambda_n) = \mathrm{GIC}^*_{\kappa_n}(\alpha_0)\right\} = P\left\{\frac{1}{n}\left(D(\lambda) - D^*(\alpha_0)\right) + \frac{\kappa_n}{n}(df_{\lambda_n} - d_{\alpha_0}) = 0\right\} \to 1. \tag{3.29}$$

The proof of (3.24) is complete.

We next show that, for any λ which cannot identify the true model, the resulting $\mathrm{GIC}_{\kappa_n}(\lambda)$ is consistently larger than $\mathrm{GIC}_{\kappa_n}(\lambda_n)$. To this end, we consider two cases, underfitting and overfitting.

Case 1: Underfitted model (i.e., $\lambda \in \Omega_-$ so that $\alpha_\lambda \not\supset \alpha_0$). Note that (3.19) and (3.24) imply that with probability tending to 1,

$$\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n) > \frac{1}{n}D^*(\alpha_\lambda) - \frac{1}{n}D^*(\alpha_0) - \frac{\kappa_n}{n}d_{\alpha_0}. \tag{3.30}$$

Taking $\inf_{\lambda\in\Omega_-}$ at both sides, and noting that $\alpha_\lambda \not\supset \alpha_0$,

$$\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n) > \min_{\alpha\not\supset\alpha_0} \frac{1}{n}D^*(\alpha) - \frac{1}{n}D^*(\alpha_0) - \frac{\kappa_n}{n}d_{\alpha_0}. \tag{3.31}$$

Under condition (C4) and noting that the number of underfitting models is finite,

$$P\left\{\min_{\alpha\not\supset\alpha_0} \frac{1}{n}D^*(\alpha) - \frac{1}{n}D^*(\alpha_0) - \frac{\kappa_n}{n}d_{\alpha_0} > 0\right\} = P\left\{\min_{\alpha\not\supset\alpha_0} c_\alpha - c_{\alpha_0} + o_P(1) > 0\right\} \to 1, \tag{3.32}$$

as $n \to \infty$. As a result,

$$P\left\{\inf_{\lambda\in\Omega_-} \mathrm{GIC}_{\kappa_n}(\lambda) > \mathrm{GIC}_{\kappa_n}(\lambda_n)\right\} \to 1. \tag{3.33}$$

Case 2: Overfitted model (i.e., $\lambda \in \Omega_+$ so that $\alpha_\lambda \supsetneq \alpha_0$). According to Lemma 3.1, with probability tending to 1, $D(\lambda_n) = D^*(\alpha_0)$. In addition, Proposition 1 indi-

cates that $df_\lambda - df_{\lambda_n} > \xi + o_P(1)$ for some $\xi > 0$ and $\lambda \in \Omega_+$. Moreover, as noticed in the proof of Lemma 3.1, $D(\lambda) \geq D^*(\alpha_\lambda)$. Therefore, with probability tending to 1,

$$\begin{aligned} n(\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n)) &= D(\lambda) - D(\lambda_n) + (df_\lambda - df_{\lambda_n})\kappa_n \\ &\geq D^*(\alpha_\lambda) - D^*(\alpha_0) + (\xi + o_P(1))\kappa_n. \quad (3.34) \end{aligned}$$

Because for any $\alpha \supsetneqq \alpha_0$, $D^*(\alpha_0) - D^*(\alpha)$ follows a χ^2 distribution with $d_\alpha - d_{\alpha_0}$ degrees of freedom asymptotically, it is of order $O_P(1)$. Accordingly,

$$\inf_{\lambda \in \Omega_+} n(\mathrm{GIC}_{\kappa_n}(\lambda) - \mathrm{GIC}_{\kappa_n}(\lambda_n)) \geq \min_{\alpha \supsetneqq \alpha_0} \{D^*(\alpha_0) - D^*(\alpha)\} + (\xi + o_P(1))\kappa_n \approx \xi\kappa_n. \quad (3.35)$$

Using the fact that κ_n goes to infinity as $n \to \infty$, we have the right hand side goes to positive infinity which guarantees the left hand side is positive as $n \to \infty$. Hence we finally have

$$P\left\{ \inf_{\lambda \in \Omega_+} \mathrm{GIC}_{\kappa_n}(\lambda) > \mathrm{GIC}_{\kappa_n}(\lambda_n) \right\} \to 1. \quad (3.36)$$

The results of Cases 1 and 2 complete the proof of (3.25). □

Proof of Theorem 3.1.

Lemma 3.1 implies that $\mathrm{GIC}_{\kappa_n}(\lambda)$, which produces the underfitted model, is consistently larger than $\mathrm{GIC}^*_{\kappa_n}(\bar{\alpha})$. Thus, the optimal model selected by minimizing the $\mathrm{GIC}_{\kappa_n}(\lambda)$ must contain all of the significant variables with probability tending to one. In addition, Lemma 3.1 indicates that there is a nonzero probability that the smallest value of $\mathrm{GIC}_{\kappa_n}(\lambda)$ for $\lambda \in \Omega_0$ is larger than that of the full model. As a result, there is a positive probability that any λ associated with the true model

cannot be selected by $\mathrm{GIC}_{\kappa_n}(\lambda)$ as the regularization parameter. Theorem 3.1(A) follows.

Lemma 3.2 indicates that the model identified by λ_n converges to the true model as the sample size gets large. In addition, it shows that those λ's, which fail to identify the true model, cannot be selected by $\mathrm{GIC}_{\kappa_n}(\lambda)$ asymptotically. Theorem 3.1(B) follows. □

Chapter 4

Asymptotic Loss Efficiency

Under the assumption that the true model is included in a family of candidate models, we established the consistency of BIC-type selectors. In practice, however, this assumption may not be valid, and hence the discussion of consistency is irrelevant. This motivates us to study the asymptotic efficiency of the AIC-type selector. In the literature, the L_2 norm has been commonly used to assess the efficiency of the classical AIC-type selector (see Shibata, 1981, 1984 and Li, 1987) in linear regression models. Hence, we first focus on the efficiency of linear regression model selections via L_2 norm. In the end of this chapter, we briefly discuss some results of asymptotic Kullback-Leibler efficiency for generalized linear model (GLIM).

4.1 Penalized Least Squares Estimation for Linear Regression

Consider the following model

$$y_i = \mu_i + \epsilon_i, \text{ for } i = 1, ..., n, \tag{4.1}$$

where $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$ is an unknown mean vector, and ϵ_i's are independent and identically distributed (i.i.d) random errors with mean 0 and variance σ^2. Furthermore, we assume that $\mathbf{X}\boldsymbol{\beta}$ constitutes the nearest representation of the true mean vector $\boldsymbol{\mu}$, and hence the full model is not necessary a correct model. Adapting the formulation of Li (1987), we allow the dimension of $\boldsymbol{\beta}$ to tend to infinity with n. Again, the goal of variable selection is to select a best model α in a candidate set $\mathcal{A}$ that gives the best approximation.

For the given data set $\{(x_i, y_i) : i = 1, \cdots, n\}$, we follow the formulation of (3.11) to define the penalized least squares function

$$Q^{LS}(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|). \tag{4.2}$$

The resulting penalized estimate of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$. In addition, the non-penalized estimate of $\boldsymbol{\mu}$ in model α is $\hat{\boldsymbol{\mu}}_\alpha^* = \mathbf{X}\hat{\boldsymbol{\beta}}_\alpha^*$. In practice, the tuning parameter λ determines the property of the penalized estimate. Adopting the GIC tuning parameter selector introduced in chapter 3, this regularization parameter can be selected by minimizing

$$\text{GIC}_{\kappa_n}^{LS}(\lambda) = \frac{1}{n}\left\{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\lambda)^2 + \kappa_n\sigma^2 d_{\alpha_\lambda}\right\}. \tag{4.3}$$

4.1.1 L_2 Loss and Risk

To assess the performance of an estimate, we adopt the approach of Shibata (1981) (also see Li, 1987; Shao, 1997), and define the average squared loss (or the L_2 loss) as follows.

Definition 4.1 (L_2 Loss and Risk). *Let $\hat{\boldsymbol{\beta}}$ be an estimate of $\boldsymbol{\beta}$ (either a penalized or unpenalized estimate). The average squared loss or the L_2 loss associated it is defined as*

$$L(\hat{\boldsymbol{\beta}}) = \frac{1}{n} \parallel \boldsymbol{\mu} - \hat{\boldsymbol{\mu}} \parallel^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\mu_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right)^2. \tag{4.4}$$

Accordingly, the L_2 risk is $R(\hat{\boldsymbol{\beta}}) = E\left[L(\hat{\boldsymbol{\beta}}) \right]$.

We first study the L_2 loss of unpenalized least squares estimate (LSE). Under model α, the LSE is the projection of $\mathbf{y}$ onto the space spanned by $\mathbf{X}_\alpha$. That is

$$\hat{\boldsymbol{\mu}}_\alpha = \mathbf{X}_\alpha \hat{\boldsymbol{\beta}}_\alpha^* = \mathbf{H}_\alpha \mathbf{y}, \tag{4.5}$$

where $\mathbf{H}_\alpha = \mathbf{X}_\alpha (\mathbf{X}_\alpha^T \mathbf{X}_\alpha)^{-1} \mathbf{X}_\alpha^T$ is the projection matrix. By some simple algebra, the squared loss (4.4) of a LSE $\hat{\boldsymbol{\beta}}_\alpha^*$ can be written as

$$L(\hat{\boldsymbol{\beta}}_\alpha^*) = \Delta_\alpha + \frac{\boldsymbol{\epsilon}^T \mathbf{H}_\alpha \boldsymbol{\epsilon}}{n}, \tag{4.6}$$

where $\Delta_\alpha = (||\boldsymbol{\mu} - \mathbf{H}_\alpha \boldsymbol{\mu}||^2)/n$ is the distance from the projection of $\boldsymbol{\mu}$ to itself. When the model α is correct, $\Delta_\alpha = 0$ and

$$L(\hat{\boldsymbol{\beta}}_\alpha^*) = \frac{\boldsymbol{\epsilon}^T \mathbf{H}_\alpha \boldsymbol{\epsilon}}{n}. \tag{4.7}$$

Note that $E[\boldsymbol{\epsilon}^T \mathbf{H}_\alpha \boldsymbol{\epsilon}] = tr(\mathbf{H}_\alpha E(\boldsymbol{\epsilon}^T \boldsymbol{\epsilon})) = \sigma^2 d_\alpha$, so we have

$$R(\hat{\boldsymbol{\beta}}_\alpha^*) = \Delta_\alpha + \frac{\sigma^2 d_\alpha}{n}, \tag{4.8}$$

where $d_\alpha = tr(\mathbf{H}_\alpha)$ is the dimension of the model. Therefore we know that in order to find a model with smaller risk, we should either find a good projection (i.e. do not underfit) or reduce the dimension of the model.

The residual sum of squares (RSS) of a LSE under model α is defined as

$$RSS(\alpha) = ||\mathbf{y} - \hat{\boldsymbol{\mu}}_\alpha||^2. \tag{4.9}$$

We have

$$\frac{RSS(\alpha)}{n} = \frac{\boldsymbol{\epsilon}^T (\mathbf{I} - \mathbf{H}_\alpha)\boldsymbol{\epsilon}}{n} + \Delta_\alpha + \frac{2\boldsymbol{\epsilon}^T (\mathbf{I} - \mathbf{H}_\alpha)\boldsymbol{\mu}}{n}. \tag{4.10}$$

When the model α is correct, $\Delta_\alpha = 0$ and the last part above is zero because $\boldsymbol{\mu} = \mathbf{X}_\alpha \boldsymbol{\beta}_\alpha$ is in the space spanned by $\mathbf{X}_\alpha$. In this case, the formula becomes

$$\frac{RSS(\alpha)}{n} = \frac{||\boldsymbol{\epsilon}||^2}{n} - L(\hat{\boldsymbol{\beta}}_\alpha^*) \left(= \frac{\boldsymbol{\epsilon}^T (\mathbf{I} - \mathbf{H}_\alpha)\boldsymbol{\epsilon}}{n} \right). \tag{4.11}$$

The first term is random noise that we cannot control. Therefore this is in line with the understanding that overfitting (resulting in a smaller RSS) will actually enlarge the loss!

Furthermore, note that

$$E\left[\frac{RSS(\alpha)}{n}\right] = \sigma^2 + \Delta_\alpha - \frac{\sigma^2 d_\alpha}{n}. \tag{4.12}$$

Comparing with (4.8), it seems that we should use

$$C_p(\alpha) = \frac{RSS(\alpha)}{n} + \frac{2\sigma^2 d_\alpha}{n}, \tag{4.13}$$

as an unbiased estimate of the risk. This was Mallows's heuristic to define C_p (Mallows, 1973).

Next we turn to the penalized estimate. Note that the LSE $\hat{\boldsymbol{\mu}}_\alpha$ is the projection of $\mathbf{y}$ onto $span(\mathbf{X}_\alpha)$. The PLSE $\hat{\boldsymbol{\mu}}_\lambda$ with a resulting model α_λ also belongs to a subspace $span(\mathbf{X}_{\alpha_\lambda})$. However, $\hat{\boldsymbol{\mu}}_\lambda$ is no longer the projection of $\mathbf{y}$ onto this space. Therefore we have the following decomposition.

$$RSS(\lambda) = ||\mathbf{y} - \hat{\boldsymbol{\mu}}_\lambda||^2 = RSS(\alpha_\lambda) + ||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2. \tag{4.14}$$

The cross term here disappears due to the orthogonality of $\mathbf{y} - \hat{\boldsymbol{\mu}}_{\alpha_\lambda}$ and $span(\mathbf{X}_{\alpha_\lambda})$. However, $\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_{\alpha_\lambda}$ is not orthogonal to $span(\mathbf{X}_{\alpha_\lambda})$, and therefore

$$L(\lambda) = L(\alpha_\lambda) + \frac{1}{n}||\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda||^2 - \frac{2}{n}\boldsymbol{\epsilon}^T(\hat{\boldsymbol{\mu}}_{\alpha_\lambda} - \hat{\boldsymbol{\mu}}_\lambda). \tag{4.15}$$

Note that when the tuning parameter $\lambda = 0$, the size of regularization is zero and hence PLSE becomes LSE. The difference between LSE and PLSE should be very small when λ is small. Intuitively, if this difference is negligible compared to the loss $L(\alpha_\lambda)$, the asymptotic properties of PLSE should be similar to those of LSE.

4.2 Asymptotic Loss Efficiency

4.2.1 Definition and Technical Conditions

Using the average squared loss measure, we further define the asymptotic loss efficiency as follows.

Definition 4.2 (Asymptotic Loss Efficiency). *A tuning parameter selection procedure is said to be asymptotically loss efficient if*

$$\frac{L(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})}{\inf_{\lambda\in[0,\lambda_{\max}]} L(\hat{\boldsymbol{\beta}}_{\lambda})} \to 1, \tag{4.16}$$

in probability, where $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ is associated with the tuning parameter $\hat{\lambda}$ selected by this procedure. We also say $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ is asymptotically loss efficient if (4.16) holds.

Moreover, we introduce the following technical conditions for studying the asymptotic loss efficiency of the $\mathrm{GIC}^{LS}_{\kappa_n}$ selector (4.3).

(E1) $(\frac{1}{n}\mathbf{X}^T\mathbf{X})^{-1}$ exists, and its largest eigenvalue is bounded by a constant number C.

(E2) $E\epsilon_1^{4q} < \infty$, for some positive integer q.

(E3) The risks of the least squares estimates $\hat{\boldsymbol{\beta}}^*_{\alpha}$ for all $\alpha \in \mathcal{A}$ satisfy

$$\sum_{\alpha\in\mathcal{A}} [nR(\hat{\boldsymbol{\beta}}^*_{\alpha})]^{-q} \to 0. \tag{4.17}$$

(E4) Let $\mathbf{b} = (b_1, ..., b_d)^T$, where $b_j = p'_{\lambda}(|\hat{\beta}_{\lambda j}|)\mathrm{sgn}(\hat{\beta}_{\lambda j})$ for all j such that $|\hat{\beta}_{\lambda j}| > 0$, and $b_j = 0$ otherwise, and $\hat{\beta}_{\lambda j}$ is the j-th component of the penalized estimate $\hat{\boldsymbol{\beta}}_{\lambda}$. In addition, let $\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}$ be the least squares estimate of β obtained from

model α_λ. Then, we assume that, in probability,

$$\sup_{\lambda\in[0,\lambda_{\max}]} \frac{\| \mathbf{b} \|^2}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \to 0. \tag{4.18}$$

Condition (E1) has been commonly considered in the literature. Conditions (E2) and (E3) are adopted from conditions (A.2) and (A.3), respectively, in Li (1987). It can be shown that if the true model is approximated by a set of the candidate models (e.g., the true model is of infinite dimension), then Condition (E3) holds. Condition (E4) ensures that the difference between the penalized mean function estimate and the corresponding least squares mean function estimate is small in comparison with the risk of the least squares estimate (see Lemma 3). Sufficient conditions for (E4) are also given in section 4.2.3. We next show the asymptotic efficiency of GIC.

4.2.2 Asymptotic Loss Efficiency of GIC selector

Theorem 4.1. *Assume conditions (E1)—(E4) hold. Then, the tuning parameter* $\hat{\lambda}$ *selected by minimizing* $GIC^{LS}_{\kappa_n}(\lambda)$ *in (4.3) with* $\kappa_n = 2$ *yields an asymptotically loss efficient estimate,* $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$*, in the sense of (4.16).*

Remark 4.1. Theorem 4.1 demonstrates that the C_p tuning parameter selector is asymptotically loss efficient. In addition, using the result that $\log(1+t) \approx t$ for small t, the C_p selector behaves similarly to the following AIC selector,

$$\mathrm{AIC}^*(\lambda) = \log(\hat{\sigma}^2_\lambda) + \frac{2\sigma^2 d_{\alpha_\lambda}}{n}. \tag{4.19}$$

Accordingly, the AIC-type selector is asymptotically loss efficient.

Remark 4.2. Applying Lemma 4.3 and Equation (4.35), we find that if

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{(\kappa_n-2)\sigma^2 d_{\alpha_\lambda}}{nR(\hat{\boldsymbol{\beta}}_\lambda)}\right|\to 0, \tag{4.20}$$

in probability, then $\mathrm{GIC}_{\kappa_n}^{LS}$ is asymptotically loss efficient. However, this is not true in general. Therefore, similarly as for classical variable selection criteria, $\kappa_n = 2$ is crucial in establishing the asymptotic loss efficiency (see Shibata, 1980, Shao, 1997 and Yang, 2005). Other GIC tuning parameter selectors with fixed $\kappa_n \neq 2$ are not necessarily asymptotically loss efficient. Specifically, BIC-type tuning parameter selectors with $\kappa_n = \log(n)$ normally do not possess asymptotic loss efficiency, which is consistent with the finding of classical variable selections (see Li, 1987 and Shao, 1997).

Before proving the theorem, we establish the following three lemmas. Lemma 4.1 means that we only need to show a variable selection criterion is asymptotically (uniformly) equivalent to the loss to establish the loss efficiency. It is applicable to both classical and modern continuously penalized methods. For the sake of simplicity, we state the Lemma with λ but keep in mind that it still holds if they are replaced by α. Lemma 4.2 evaluates the difference between a penalized mean estimate $\hat{\boldsymbol{\mu}}_\lambda$ and its corresponding least squares mean estimate $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}$, while Lemma 4.3 demonstrates that the losses of $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}$ are asymptotically equivalent.

Lemma 4.1. *Suppose $L(\lambda) > 0$ for all $\lambda \in \Lambda$. Assume $C(\lambda) = L(\lambda) + r(\lambda)$, where*

$$\sup_{\lambda\in\Lambda}\left|\frac{r(\lambda)}{L(\lambda)}\right|\to 0, \quad \textit{in probability.} \tag{4.21}$$

Let $\hat{\lambda}_n = arginf_{\lambda\in\Lambda} C(\lambda)$. *Assume* $L(\hat{\lambda}_n)$ *is bounded. We have*

$$\frac{L(\hat{\lambda}_n)}{\inf_{\lambda\in\Lambda} L(\lambda)} \to 1, \quad \textit{in probability.} \tag{4.22}$$

Proof. With probability tending to 1,

$$L(\hat{\lambda}_n) = C(\hat{\lambda}_n) - \frac{r(\hat{\lambda}_n)}{L(\hat{\lambda}_n)} L(\hat{\lambda}_n) \leq C(\lambda) + \left|\frac{r(\hat{\lambda}_n)}{L(\hat{\lambda}_n)}\right| L(\hat{\lambda}_n), \text{ for any } \lambda \in \Lambda. \tag{4.23}$$

Take $\inf_{\lambda\in\Lambda}$ at the right hand side, we have

$$\begin{aligned} \inf_{\lambda\in\Lambda} L(\lambda) \leq L(\hat{\lambda}_n) &\leq \inf_{\lambda\in\Lambda} \left\{ L(\lambda) \left[1 + \frac{r(\lambda)}{L(\lambda)}\right] \right\} + \frac{r(\hat{\lambda}_n)}{L(\hat{\lambda}_n)} L(\hat{\lambda}_n) && (4.24) \\ &\leq \inf_{\lambda\in\Lambda} L(\lambda) \left[1 + \sup_{\lambda\in\Lambda} \left|\frac{r(\lambda)}{L(\lambda)}\right|\right] + \left|\frac{r(\hat{\lambda}_n)}{L(\hat{\lambda}_n)}\right| L(\hat{\lambda}_n), && (4.25) \end{aligned}$$

where the right hand side goes to $\inf_{\lambda\in\Lambda} L(\lambda)$ as n goes to ∞, given the conditions in the lemma. The lemma therefore holds. □

Lemma 4.2. *Under condition (E1),*

$$\| \hat{\boldsymbol{\mu}}_\lambda - \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} \|^2 \leq nC \| \mathbf{b} \|^2, \tag{4.26}$$

where C *is the constant number in condition (E1) and* $\mathbf{b}$ *is defined in condition (E4).*

Proof. Without loss of generality, we assume that the first d_{α_λ} components of $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}$ are nonzero, and denote them by $\hat{\boldsymbol{\beta}}^{(1)}_\lambda$ and $\hat{\boldsymbol{\beta}}^{*(1)}_{\alpha_\lambda}$, respectively. Thus, $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda = \mathbf{X}_{\alpha_\lambda}\hat{\boldsymbol{\beta}}^{(1)}_\lambda$ and $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda} = \mathbf{X}\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda} = \mathbf{X}_{\alpha_\lambda}\hat{\boldsymbol{\beta}}^{*(1)}_{\alpha_\lambda}$. From the proofs of Theorems 1 and 2 in Fan and Li (2001), with probability tending to 1, we have that $\hat{\boldsymbol{\beta}}^{(1)}_\lambda$ is

the solution of the following equation,

$$\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^T\left(\mathbf{y}-\mathbf{X}_{\alpha_\lambda}\boldsymbol{\beta}_\lambda^{(1)}\right)+\mathbf{b}^{(1)}=\mathbf{0}, \tag{4.27}$$

where $\mathbf{b}^{(1)}$ is the subvector of $\mathbf{b}$ that corresponds to $\hat{\boldsymbol{\beta}}_\lambda^{(1)}$. Accordingly,

$$\hat{\boldsymbol{\beta}}_\lambda^{(1)}=\left(\mathbf{X}_{\alpha_\lambda}^T\mathbf{X}_{\alpha_\lambda}\right)^{-1}\mathbf{X}_{\alpha_\lambda}^T\mathbf{y}+\left(\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^T\mathbf{X}_{\alpha_\lambda}\right)^{-1}\mathbf{b}^{(1)}=\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)}+\mathbf{V}_{\alpha_\lambda}\mathbf{b}^{(1)}, \tag{4.28}$$

where $\mathbf{V}_{\alpha_\lambda}\triangleq\left(\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^T\mathbf{X}_{\alpha_\lambda}\right)^{-1}$. In addition, the eigenvalues of $\mathbf{V}_{\alpha_\lambda}$ are bounded under condition (E1). Hence,

$$\parallel\hat{\boldsymbol{\mu}}_\lambda-\hat{\boldsymbol{\mu}}_{\alpha_\lambda}^*\parallel^2=\parallel\mathbf{X}_{\alpha_\lambda}(\hat{\boldsymbol{\beta}}_\lambda^{(1)}-\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)})\parallel^2=n\mathbf{b}_1^T\mathbf{V}_{\alpha_\lambda}\mathbf{b}^{(1)}\leq nC\parallel\mathbf{b}\parallel^2, \tag{4.29}$$

for some positive constant number C. This completes the proof. □

Lemma 4.3. *If conditions (E1)—(E4) hold, then*

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^*)}-1\right|\to 0, \tag{4.30}$$

in probability.

Proof. After algebraic simplification, we have

$$L(\hat{\boldsymbol{\beta}}_\lambda)-L(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^*)=\frac{\parallel\hat{\boldsymbol{\mu}}_{\alpha_\lambda}^*-\hat{\boldsymbol{\mu}}_\lambda\parallel^2}{n}+\frac{2(\boldsymbol{\mu}-\hat{\boldsymbol{\mu}}_{\alpha_\lambda}^*)^T(\hat{\boldsymbol{\mu}}_{\alpha_\lambda}^*-\hat{\boldsymbol{\mu}}_\lambda)}{n}=I_1+I_2. \tag{4.31}$$

Under conditions (E2) and (E3), Li (1987) showed that

$$\sup_{\alpha\in\mathcal{A}}\left|\frac{L(\hat{\boldsymbol{\beta}}_\alpha^*)}{R(\hat{\boldsymbol{\beta}}_\alpha^*)}-1\right|\to 0. \tag{4.32}$$

This, together with condition (E4) and Lemma 4.2, implies

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{I_1}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\right| = \sup_{\lambda\in[0,\lambda_{\max}]}\left\{\frac{\parallel \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda\parallel^2}{nR(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}-\frac{\parallel \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda\parallel^2}{nL(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\left[\frac{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}-1\right]\right\}\to 0. \tag{4.33}$$

Applying the Cauchy-Schwarz inequality, we next obtain

$$I_2 \le \frac{2\parallel \boldsymbol{\mu}-\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}\parallel\cdot\parallel \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda\parallel}{n} = 2\sqrt{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\cdot\frac{1}{\sqrt{n}}\parallel \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda\parallel. \tag{4.34}$$

As a result, $\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{I_2}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\right|\to 0$, and Lemma 4.3 follows immediately. □

Proof of Theorem 4.1. To show the asymptotic efficiency of $\mathrm{GIC}_2^{LS}(\lambda)$, from Lemma 4.1, it suffices to demonstrate that minimizing $\mathrm{GIC}_2^{LS}(\lambda)$ is the same as minimizing $L(\hat{\boldsymbol{\beta}}_\lambda)$ asymptotically. To this end, we need to prove that, in probability,

$$\sup_{\lambda\in[0,\lambda_{\max}]}\left|\frac{\mathrm{GIC}_2^{LS}(\lambda)-\frac{\|\boldsymbol{\epsilon}\|^2}{n}-L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}_\lambda)}\right|\to 0. \tag{4.35}$$

Let the projection matrix corresponding to the model α be $\mathbf{H}_\alpha=\mathbf{X}_\alpha(\mathbf{X}_\alpha^T\mathbf{X}_\alpha)^{-1}\mathbf{X}_\alpha^T$. Then,

$$\begin{aligned}
\mathrm{GIC}_2^{LS}(\lambda) &= \frac{\parallel \mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda\parallel^2}{n}+\frac{2\sigma^2 d_{\alpha_\lambda}}{n}\\
&= \frac{\parallel \mathbf{y}-\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}\parallel^2}{n}+\frac{\parallel \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda\parallel^2}{n}+\frac{2\sigma^2 d_{\alpha_\lambda}}{n}\\
&= \frac{\parallel \boldsymbol{\epsilon}\parallel^2}{n}+L(\hat{\boldsymbol{\beta}}_\lambda)+\left[L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})-L(\hat{\boldsymbol{\beta}}_\lambda)\right]+\frac{1}{n}\parallel \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda\parallel^2\\
&\quad +\frac{2}{n}\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{H}_{\alpha_\lambda})\boldsymbol{\mu}+\frac{2}{n}(\sigma^2 d_{\alpha_\lambda}-\boldsymbol{\epsilon}^T\mathbf{H}_{\alpha_\lambda}\boldsymbol{\epsilon}).
\end{aligned} \tag{4.36}$$

Let $J_1 = L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})-L(\hat{\boldsymbol{\beta}}_\lambda)$, $J_2 =\parallel \hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}-\hat{\boldsymbol{\mu}}_\lambda\parallel^2/n$, $J_3 = 2\boldsymbol{\epsilon}^T(\mathbf{I}-\mathbf{H}_{\alpha_\lambda})\boldsymbol{\mu}/n$, and

$J_4 = 2(\sigma^2 d_{\alpha_\lambda} - \boldsymbol{\epsilon}^T \mathbf{H}_{\alpha_\lambda} \boldsymbol{\epsilon})/n$. Using Lemma 4.2, Lemma 4.3 and similar arguments used in Li (1987), we obtain that, in probability,

$$\sup_{\lambda \in [0, \lambda_{\max}]} \left| \frac{J_j}{L(\hat{\boldsymbol{\beta}}_\lambda)} \right| \to 0, \quad \text{for } j = 1, ..., 4. \tag{4.37}$$

Accordingly, (4.35) holds, which implies that the difference between $\mathrm{GIC}_2^{LS}(\lambda) - \frac{\|\boldsymbol{\epsilon}\|^2}{n}$ and $L(\hat{\boldsymbol{\beta}}_\lambda)$ is negligible in comparison to $L(\hat{\boldsymbol{\beta}}_\lambda)$. This completes the proof. □

Remark 4.3. In practice, σ^2 is often unknown. It is natural to replace the σ^2 in GIC_2^{LS} by its consistent estimate (see Shao, 1997). The following corollary shows that the asymptotical property of GIC still holds.

Corollary 4.1. *If the tuning parameter $\hat{\lambda}$ is selected by minimizing $\mathrm{GIC}_2^{LS}(\lambda)$ with σ^2 being replaced by its consistent estimate $\tilde{\sigma}^2$, then GIC_2^{LS} is asymptotically loss efficient.*

Proof. When σ^2 is unknown, the $\mathrm{GIC}_2^{LS}(\lambda)$ in (4.36) becomes

$$\mathrm{GIC}_2^{LS}(\lambda) = \frac{\| \boldsymbol{\epsilon} \|^2}{n} + L(\hat{\boldsymbol{\beta}}_\lambda) + J_1 + J_2 + J_3 + J_4 + \frac{2(\tilde{\sigma}^2 - \sigma^2) d_{\alpha_\lambda}}{n}. \tag{4.38}$$

Using Lemma 4.3 and condition (E3), we have

$$\sup_{\lambda \in [0, \lambda_{\max}]} \left| \frac{2(\tilde{\sigma}^2 - \sigma^2) d_{\alpha_\lambda}}{n L(\hat{\boldsymbol{\beta}}_\lambda)} \right| \to 0 \tag{4.39}$$

in probability, which completes the proof. □

4.2.3 Sufficient Conditions for Condition (E4)

First we list some technical conditions that will be used in the following discussions.

(S1) There exists a constant M_1 such that $\lambda_{\max}$ satisfies $\sqrt{n}\lambda_{\max} < M_1$ for all n.

(S2) There exists a constant M_2 such that $p_\lambda(\theta)$ satisfies $p'_\lambda(\theta) \leq M_2\lambda$ for any θ.

We provide motivations for these conditions. Assume that the true model is $\mu_i = \mathbf{x}_i^T\boldsymbol{\beta}_0 = \sum_{j=1}^d x_{ij}\beta_{j0}$ for $i = 1, \cdots, n$, and then write $a_n = \max_{1\leq j\leq d}\{p'_{\lambda_n}(|\beta_{j0}|), \beta_{j0} \neq 0\}$. Under the condition $d = O(n^\nu)$ with $\nu < \frac{1}{4}$, Fan and Peng (2004) proved that there exists a local maximum of the penalized least squares function, so that $\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \| = O_P\{\sqrt{d}(n^{-1/2} + a_n)\}$. Thus, conditions (S1) and (S2) ensure that the penalized estimator is $\sqrt{n/d}$-consistent. Two sufficient conditions for (E4) are then given in the following propositions.

Proposition 4.1. *Assume that (S1) and (S2) are satisfied. Condition (E4) holds if the average error of the full model $\bar{\alpha}$, i.e. $\Delta_{\bar{\alpha}} \triangleq \| \boldsymbol{\mu} - \mathbf{H}_{\bar{\alpha}}\boldsymbol{\mu} \|^2 /n$, satisfies*

$$\frac{n\Delta_{\bar{\alpha}}}{d} \to \infty, \tag{4.40}$$

as $n \to \infty$.

Proposition 4.2. *Assume that (S1) and (S2) are satisfied. Condition (E4) holds if the penalty function further satisfies condition (C2), and that*

$$\sup_{\lambda\in[0,\lambda_{\max}]} \frac{1}{d_{\alpha_\lambda}} \sum_{j=1}^{d} I(0 < |\hat{\beta}_{\lambda j}| \leq m\lambda) \to 0, \quad \text{as } n \to \infty. \tag{4.41}$$

where m is defined in (C2).

Before we outline the technical proofs, we first give some discussions of the two propositions. In Proposition 4.1, the penalty function is only assumed to satisfy condition (S2). This includes LASSO as an example, where the bias of the penalized estimate is a great concern. Due to this systematic bias from the

penalty, condition (E4) is only satisfied when the model error is the dominant part in $R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})$. Proposition 4.1 means that the bias can be well controlled when the average model error of the full model satisfies (4.40), i.e., the rate at which $\Delta_{\bar{\alpha}}$ goes to zero cannot be faster than d/n. In other words, there is a discrepancy between the full model and the truth in the above sense.

In Proposition 4.2, no assumption is imposed on the model misspecification, but we further assume the penalty function satisfies (C2), which includes SCAD and MCP, among others. These penalty functions do not penalize large coefficients and hence result in penalized estimates with oracle properties (Fan & Li, 2001). In this situation, (4.41) means that the proportion of small or moderate size coefficients ($0 < \sqrt{n}|\hat{\beta}_{\lambda j}| \leq mM_1$) is vanishing (note that $\sqrt{n}\lambda$ is bounded by M_1). This condition (4.41) is satisfied automatically under the identifiability assumption in Fan and Peng (2004) that $\min_{j:\beta_{0j}\neq 0} |\beta_{0j}|/\lambda \to \infty$, as $n \to \infty$. As a byproduct, a similar condition can also be established for the adaptive LASSO, which employs data driven penalties with $b_j = \lambda/(\hat{\beta}^*_{\bar{\alpha}j})^k$ for some $k > 0$ to cope with the bias of LASSO. Similarly to Proposition 4.2, we can show that condition (E4) holds if the condition (4.41) is replaced by

$$\sup_{\lambda\in[0,\lambda_{\max}]} \frac{1}{d_{\alpha_\lambda}} \sum_{j=1}^{d} \frac{1}{(\hat{\beta}^*_{\bar{\alpha}j})^{2k}} I(|\hat{\beta}_{\lambda j}| > 0) \to 0, \quad \text{as } n \to \infty. \tag{4.42}$$

Proof of Proposition 4.1. First we note that

$$R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda}) = \Delta_{\alpha_\lambda} + \frac{d_{\alpha_\lambda}\sigma^2}{n} \geq \Delta_{\bar{\alpha}} + \frac{d_{\alpha_\lambda}\sigma^2}{n}. \tag{4.43}$$

Under condition (4.40), the first term in (4.43) dominates the second term. From (S2) and (4.43), we have that

$$\frac{\| \mathbf{b} \|^2}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \leq \frac{M_2^2 \lambda^2 d_{\alpha_\lambda}}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \leq M_2^2 \cdot n\lambda_{\max}^2 \cdot \frac{1}{n\Delta_{\bar{\alpha}}/d}, \tag{4.44}$$

which goes to zero independently of λ. Hence condition (E4) holds.

Proof of Proposition 4.2. Under condition (C2), the components in $\mathbf{b}$ are zero except for those $\hat{\beta}_{\lambda j} \leq m\lambda$. Similarly, we use (S2) and (4.43) and conclude that

$$\frac{\| \mathbf{b} \|^2}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \leq \frac{\| \mathbf{b} \|^2}{n^{-1} d_{\alpha_\lambda} \sigma^2} \leq \frac{M_2^2}{\sigma^2} \cdot n\lambda_{\max}^2 \cdot \frac{1}{d_{\alpha_\lambda}} \sum_{j=1}^{d} I(0 < |\hat{\beta}_{\lambda j}| \leq m\lambda), \tag{4.45}$$

where the first term is constant, the second term is bounded under (S1) and the third term goes to zero uniformly in λ under (4.41).

4.3 Generalized Linear Model

In this section, we study the asymptotic loss efficiency for the generalized linear model (GLIM) under Kullback-Leibler (KL) loss. It is shown later that the KL loss is identical to the squared loss for the normal linear regression model. Furthermore, under mild conditions, we demonstrate that the KL loss is asymptotically equivalent to a weighted squared loss taking into account the heteroscedastic error.

To the best of our knowledge, there is no work on asymptotic loss efficiency for GLIM. Extension of asymptotic loss efficiency from the linear regression model to GLIM indeed is very challenging in that the Taylor expansions, the main mathematical tool in the asymptotic analysis, cannot be utilized when a candidate model is not in the neighborhood of the true model. This requires us to establish a general

framework to develop the theory. To this end, (a) we first present the asymptotic theory for GLIM when the working model may be only an approximation of the true one. The asymptotic theory allows us to analyze the asymptotic bias and variance of the resulting estimate. (b) From the asymptotic bias, we found that the resulting estimate may not be consistent when the working model is misspecified. Thus, we restrict ourselves to candidate models which are in the neighborhood of the true model, so that Taylor expansions are allowed. (c) We established the KL asymptotic loss efficiency with the classical AIC best subset selection, which is already new in the literature. (d) We finally show the KL loss efficiency of the penalized likelihood estimate with the AIC tuning parameter selector.

4.3.1 Asymptotic Theory of GLIM Estimate

It is well known that when the working model is a correct model, under certain regularity conditions, the resulting estimate of the GLIM is consistent and follows an asymptotic normal distribution. In this section, we present the asymptotic theory of GLIM estimate without assuming that the working model is a correct model. Consider that a sample of size n is collected from the generalized linear model, whose density function is

$$f(y;\theta,\phi) = \exp\left\{ \left[y\theta - b(\theta)\right]/a(\phi) + c(y,\phi)\right\}, \tag{4.46}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot,\cdot)$ are suitably chosen functions, θ is the canonical parameter, and ϕ is a scale parameter which is also called the dispersion parameter. Additionally, $a(\cdot)$ is assumed to be positive and $b(\cdot)$ is a second order smooth function with $b''(\theta) > 0$ and $b'''(\theta)$ bounded for every $\theta \in \Theta$. Given θ, denote the mean and variance of y by μ and σ^2. It can be easily shown that $E(y) = \mu = b'(\theta)$ and

$Var(y) = \sigma^2 = a(\phi)b''(\theta)$. In general, the canonical parameter θ is related with the systematic parameter η_i though the pre-specified link function $g(\cdot)$, so that $\eta = g(\mu) = g \circ b'(\theta)$. In this paper, we restrict the discussion to canonical link functions. That is, we assume $g^{-1}(\cdot) = b'(\cdot)$ so that $\theta = \eta$. Furthermore, denote the true canonical parameter by $\boldsymbol{\theta}_0 = (\theta_{01}, ..., \theta_{0n})^T$. For all i, we assume that θ_{0i} lies in a set Θ with bounded support, and $\boldsymbol{\theta}_0$ is an interior point of Θ^n.

Suppose the canonical parameter θ is a function of covariates $\mathbf{x}$. Write $\theta = \theta(\mathbf{x})$. In practice, some d-dimensional model $\bar{\alpha}$ along with some submodels $\alpha \subset \bar{\alpha}$ are considered. The goal of variable selection is to find a best submodel that is parsimonious while approximating $\theta(\mathbf{x})$ well. For the sake of simplicity, we only consider the case where the dispersion parameter ϕ is known. This includes the normal linear regression model with known variance, the logistic regression model and the Poisson log-linear model. Our results continue to hold when a consistent estimate of the dispersion parameter is available, for example, when it can be estimated consistently from the full model. This is common in the context of variable selection.

Throughout the paper we consider a fixed design where $\mathbf{x}_i$, $i = 1, ..., n$ are assumed to be non-random. The results in this paper are also valid in the almost sure sense when the $\mathbf{x}_i$'s are random, provided that the required conditions involving $\mathbf{x}_i$'s hold for $n = 1, 2, ...$ This is similar to the case for classical variable selection criteria (Shao, 1997).

Using the likelihood principle, the log-likelihood function under model α can be written as

$$\ell(\boldsymbol{\beta}_\alpha) = \sum_{i=1}^{n} \left[\frac{y_i \mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha - b(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha)}{a(\phi)} + c(y_i, \phi) \right], \tag{4.47}$$

where $\boldsymbol{\beta}_\alpha$ can be estimated by the maximum likelihood estimate (MLE), denoted

by $\hat{\boldsymbol{\beta}}_\alpha^* = \text{argmax}_{\boldsymbol{\beta}_\alpha} \ell(\boldsymbol{\beta}_\alpha)$.

Denote by f_0 the true (conditional) distribution function of $y|\mathbf{x}$, and let

$$f_\alpha(y; \boldsymbol{\beta}_\alpha, \phi) = \exp\left\{ \left[y\mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha - b(\mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha) \right] / a(\phi) + c(y, \phi) \right\}, \tag{4.48}$$

be the candidate distribution under model α with parameter $\boldsymbol{\beta}_\alpha$. The Kullback-Leibler (KL) discrepancy between the true model and the candidate model is

$$\begin{aligned} \rho\left(\theta(\mathbf{x}), \mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha\right) &= E_0 \left\{ \log\left(\frac{f_0}{f_\alpha} \right) \right\} \\ &= \frac{1}{a(\phi)} \left[b'(\theta(\mathbf{x}))(\theta(\mathbf{x}) - \mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha) - b(\theta(\mathbf{x})) + b(\mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha) \right] \end{aligned} \tag{4.49}$$

where E_0 means the expectation is taken under the true distribution f_0.

Note that the second derivative of $n^{-1} \sum_{i=1}^n \rho\left(\theta(\mathbf{x}), \mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha\right)$ with respect to $\boldsymbol{\beta}_\alpha$ is $n^{-1} \sum_{i=1}^n b''(\mathbf{x}_{\alpha i} \boldsymbol{\beta}_\alpha) \mathbf{x}_{\alpha i} \mathbf{x}_{\alpha i}^T$, which is positive definite since the variance function $b''(\mathbf{x}^T \boldsymbol{\beta}) > 0$. Thus, the objective function $n^{-1} \sum_{i=1}^n \rho\left(\theta(\mathbf{x}), \mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha\right)$ to minimize is convex. Adopting from Hjort and Pollard (1993), a key assumption we make here is that, for each candidate model α, there exists a unique optimal parameter $\boldsymbol{\beta}_\alpha^*$ which is the minimizer of the limit of $n^{-1} \sum_{i=1}^n \rho\left(\theta(\mathbf{x}_i), \mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha\right)$ over $\boldsymbol{\beta}_\alpha$. Although randomness in $\mathbf{x}$ is not necessary, it helps to understand this optimal parameter if we consider for now that the distribution of $\mathbf{x}$ is F_x, $\boldsymbol{\beta}_\alpha^*$ is actually the minimizer of the weighted KL discrepancy $\boldsymbol{\beta}_\alpha^* = \text{argmin}_{\boldsymbol{\beta}_\alpha} \int \rho\left(\theta(\mathbf{x}), \mathbf{x}_\alpha^T \boldsymbol{\beta}_\alpha\right) dF_x(\mathbf{x})$.

Next we introduce some notations following Hjort and Pollard (1993), so that we can apply their Theorem 2.3 in section 2C to establish the asymptotic distribution of $\hat{\boldsymbol{\beta}}_\alpha^*$. Write $g_i(y_i, \boldsymbol{\beta}_\alpha | \mathbf{x}_i) = a(\phi)^{-1}[-y_i \mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha + b(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha)]$, so the MLE $\hat{\boldsymbol{\beta}}_\alpha^*$

is the minimizer of $\sum_{i=1}^{n} g_i(y_i, \boldsymbol{\beta}_\alpha|\mathbf{x}_{\alpha i})$. We can write

$$\begin{aligned} g_i(y_i, \boldsymbol{\beta}_\alpha^* + \mathbf{t}|\mathbf{x}_{\alpha i}) - g_i(y_i, \boldsymbol{\beta}_\alpha^*|\mathbf{x}_{\alpha i}) &= \frac{1}{a(\phi)}\left[-y_i \mathbf{x}_{\alpha i}^T \mathbf{t} + b(\mathbf{x}_{\alpha i}^T(\boldsymbol{\beta}_\alpha^* + \mathbf{t})) - b(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*)\right] \\ &= \{\delta(\mathbf{x}_{\alpha i}) + D_i(y_i|\mathbf{x}_{\alpha i})\}^T \mathbf{t} + R_i(y_i, \mathbf{t}|\mathbf{x}_{\alpha i}), \end{aligned} \tag{4.50}$$

where

$$\delta(\mathbf{x}_{\alpha i}) = a(\phi)^{-1}\left[b'(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*) - b'(\theta(\mathbf{x}_i))\right]\mathbf{x}_{\alpha i}, \tag{4.51}$$

$$D_i(y_i|\mathbf{x}_{\alpha i}) = -a(\phi)^{-1}\left[y_i - b'(\theta(\mathbf{x}_i))\right]\mathbf{x}_{\alpha i}, \tag{4.52}$$

$$R_i(y_i, \mathbf{t}|\mathbf{x}_{\alpha i}) = a(\phi)^{-1}\left[b(\mathbf{x}_{\alpha i}^T(\boldsymbol{\beta}_\alpha^* + \mathbf{t})) - b(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*) - b'(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*)\mathbf{x}_{\alpha i}^T \mathbf{t}\right]. \tag{4.53}$$

It is easy to verify that $E(D_i(y_i|\mathbf{x}_{\alpha i})) = 0$ because $Ey_i = b'(\theta(\mathbf{x}_i))$. Denote

$$\mathbf{B}_i(\mathbf{x}_{\alpha i}) = Var(D_i(y_i|\mathbf{x}_{\alpha i})) = b''(\theta(\mathbf{x}_i))\mathbf{x}_{\alpha i}\mathbf{x}_{\alpha i}^T. \tag{4.54}$$

Write furthermore,

$$ER_i(y_i, \mathbf{t}|\mathbf{x}_{\alpha i}) = \frac{1}{2}\mathbf{t}^T \mathbf{A}_i(\mathbf{x}_{\alpha i})\mathbf{t} + v_{i,0}(\mathbf{t}|\mathbf{x}_{\alpha i}) \quad \text{and} \quad VarR_i(y_i, \mathbf{t}|\mathbf{x}_{\alpha i}) = v_i(\mathbf{t}|\mathbf{x}_i), \tag{4.55}$$

where $\mathbf{A}_i(\mathbf{x}_{\alpha i}) = b''(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*)\mathbf{x}_{\alpha i}\mathbf{x}_{\alpha i}^T / a(\phi)$ and

$$v_{i,0}(\mathbf{t}|\mathbf{x}_{\alpha i}) = \frac{1}{6a(\phi)} \sum_{j=1}^{d_\alpha} \sum_{k=1}^{d_\alpha} \sum_{l=1}^{d_\alpha} b'''(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*) x_{\alpha ij} x_{\alpha ik} x_{\alpha il} \xi_j \xi_k \xi_l, \tag{4.56}$$

with $\boldsymbol{\xi}$ lying between 0 and $\mathbf{t}$. Finally, write

$$\mathbf{J}_n = \sum_{i=1}^{n} \mathbf{A}_i(\mathbf{x}_{\alpha i}) = \frac{1}{a(\phi)} \mathbf{X}_\alpha^T \mathrm{diag}\{b''(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*)\}_{i=1}^n \mathbf{X}_\alpha, \tag{4.57}$$

$$\mathbf{K}_n = \sum_{i=1}^{n} \mathbf{B}_i(\mathbf{x}_{\alpha i}) = \frac{1}{a(\phi)} \mathbf{X}_\alpha^T \mathrm{diag}\{b''(\theta(\mathbf{x}_i))\}_{i=1}^n \mathbf{X}_\alpha, \tag{4.58}$$

$$\begin{aligned} \mathbf{L}_n &= \sum_{i=1}^{n} \delta_i(\mathbf{x}_{\alpha i}) \delta_i(\mathbf{x}_{\alpha i})^T \\ &= \frac{1}{a(\phi)} \mathbf{X}_\alpha^T \mathrm{diag} \left\{ \left[b'(\mathbf{x}_{\alpha i}^T \boldsymbol{\beta}_\alpha^*) - b'(\theta(\mathbf{x}_i)) \right]^2 \right\}_{i=1}^n \mathbf{X}_\alpha. \end{aligned} \tag{4.59}$$

Theorem 4.2. *Assume that*

$$\max_{i=1,\ldots,n} \frac{|\mathbf{x}_i|}{\sqrt{n}} \to 0 \quad \textit{for } n \to \infty. \tag{4.60}$$

Also assume $\mathbf{J}_n/n$ *is bounded away from 0, and that* $\mathbf{K}_n/n$ *and* $\mathbf{L}_n/n$ *are bounded. We have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\alpha^* - \boldsymbol{\beta}_\alpha^*) = -(\mathbf{J}_n/n)^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \delta(\mathbf{x}_{\alpha i}) + n^{-1/2} \sum_{i=1}^{n} D(y_i|\mathbf{x}_{\alpha i}) \right\} + o_P(1). \tag{4.61}$$

If furthermore $\mathbf{J}_n/n \to \mathbf{J}$, $\mathbf{K}_n/n \to \mathbf{K}$ *and* $\mathbf{L}_n/n \to \mathbf{L}$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_\alpha^* - \boldsymbol{\beta}_\alpha^*) = N\left(-\mathbf{J}^{-1} n^{-1/2} \sum_{i=1}^{n} \delta(\mathbf{x}_{\alpha i}), \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-1} \right) + o_P(1). \tag{4.62}$$

Proof. First note that $v_i(\mathbf{s}/\sqrt{n}|\mathbf{x}_i) = 0$, and

$$\sum_{i=1}^{n} v_{i,0}(\mathbf{s}/\sqrt{n}|\mathbf{x}_i) \xrightarrow{P} 0 \quad \text{for each } \mathbf{s}, \tag{4.63}$$

because we assume $b'''(\cdot)$ is bounded and $\max_i |\mathbf{x}_i|/\sqrt{n} \to 0$. Therefore we can directly apply Theorem 2.3 in Hjort and Pollard (1993) and conclude (4.61) under the condition that $\mathbf{J}_n/n$ is bounded away from 0, and that $\mathbf{K}_n/n$ and $\mathbf{L}_n/n$ are bounded. (4.62) follows immediately from (4.61). □

If we write

$$\nabla \ell(\boldsymbol{\beta}_\alpha^*) = \frac{\partial \ell(\boldsymbol{\beta}_\alpha^*)}{\partial \boldsymbol{\beta}_\alpha} = a(\phi)^{-1}\mathbf{X}_\alpha^T\{\mathbf{y} - b'(\mathbf{X}_\alpha\boldsymbol{\beta}_\alpha^*)\}, \tag{4.64}$$

$$\nabla^2 \ell(\boldsymbol{\beta}_\alpha^*) = \frac{\partial^2 \ell(\boldsymbol{\beta}_\alpha^*)}{\partial \boldsymbol{\beta}_\alpha^T \partial \boldsymbol{\beta}_\alpha} = -a(\phi)^{-1}\mathbf{X}_\alpha^T\mathbf{V}_\alpha^T\mathbf{V}_\alpha\mathbf{X}_\alpha, \tag{4.65}$$

where $\mathbf{V}_\alpha = \text{diag}\{b''(\mathbf{x}_{\alpha i}^T\boldsymbol{\beta}_\alpha^*)^{1/2}, ..., b''(\mathbf{x}_{\alpha n}^T\boldsymbol{\beta}_\alpha^*)^{1/2}\}$, Theorem 4.2 means that although the asymptotic expansion of the MLE still has the form

$$\begin{aligned} \hat{\boldsymbol{\beta}}_\alpha^* - \boldsymbol{\beta}_\alpha^* &= -\left[\nabla^2 \ell(\boldsymbol{\beta}_\alpha^*)\right]^{-1} \nabla \ell(\boldsymbol{\beta}_\alpha^*) + o_P(1/\sqrt{n}) \\ &= \left(\mathbf{X}_\alpha^T\mathbf{V}_\alpha^T\mathbf{V}_\alpha\mathbf{X}_\alpha\right)^{-1}\mathbf{X}_\alpha^T\{\mathbf{y} - b'(\mathbf{X}_\alpha\boldsymbol{\beta}_\alpha^*)\} + o_P(1/\sqrt{n}), \end{aligned} \tag{4.66}$$

the key difference from the traditional theory is that $E\nabla\ell(\boldsymbol{\beta}_\alpha^*) \neq 0$ in light of an additional bias due to $\delta(\mathbf{x}_{\alpha i})$.

4.3.2 The Set of Candidate Models

Note that the center of $\hat{\boldsymbol{\beta}}_\alpha^*$'s distribution is approximately

$$\boldsymbol{\beta}_\alpha^* - (\mathbf{J}_n/n)^{-1}n^{-1/2}\sum_{i=1}^{n}\delta(\mathbf{x}_{\alpha i}). \tag{4.67}$$

When α is not a correct model such that $\theta(\mathbf{x}) \neq \mathbf{x}_\alpha\boldsymbol{\beta}_\alpha^*$, $\hat{\boldsymbol{\beta}}_\alpha^*$ is not asymptotically unbiased for the true parameter $\boldsymbol{\beta}_0$, and is not even unbiased for the optimal parameter $\boldsymbol{\beta}_\alpha^*$. Furthermore, the asymptotic variance of $\hat{\boldsymbol{\beta}}_\alpha^*$ is also related to the extent of model misspecification which is made manifest by the component $\mathbf{L}_n$. These challenging issues make it difficult to develop asymptotic theories via Taylor expansion. Therefore, in the following discussion, we restrict our focus to those

models which are close to the correct model. Define the set of candidate models

$$\mathcal{C} = \{\alpha : \sup_{1\le i\le n} |\hat{\theta}_\alpha(\mathbf{x}_i) - \theta_0(\mathbf{x}_i)| \to 0, \text{in probability}\}, \tag{4.68}$$

where $\hat{\theta}_\alpha(\mathbf{x}_i) = \mathbf{x}_{\alpha i}^T \hat{\boldsymbol{\beta}}_\alpha$ and $\theta_0(\cdot)$ is the true canonical parameter. For models in $\mathcal{C}$, the approximation bias is small so that $\hat{\theta}(\mathbf{x}_i)$ is within a neighborhood of $\theta_0(\mathbf{x}_i)$. This allows us to use Taylor expansion for $b(\hat{\theta}(\mathbf{x}_i))$ at $\theta = \theta_0(\mathbf{x}_i)$. The candidate set $\mathcal{C}$ plays the same role as the set consisting of all possible subsets of the full model in a linear regression model because we do not need the Taylor expansion for $b(\hat{\theta}(\mathbf{x}_i))$ at $\theta = \theta_0(\mathbf{x}_i)$.

The technical condition

$$\sup_{1\le i\le n} |\hat{\theta}_\alpha(\mathbf{x}_i) - \theta_0(\mathbf{x}_i)| \to 0 \quad \text{in probability} \tag{4.69}$$

in (4.68) is critical. It restricts our focus to models in which the estimates $\hat{\boldsymbol{\theta}}_\alpha$ are close to the true parameter $\boldsymbol{\theta}_0$. Without this condition, Taylor expansion cannot be performed and hence all arguments based on it will not hold.

4.3.3 Asymptotic Representation of KL Loss

Recall that the true canonical parameter is denoted by $\boldsymbol{\theta}_0 = (\theta_{01}, ..., \theta_{0n})$. Write E_0 and Var_0 as the expectation and variance under the truth; then $\boldsymbol{\theta}_0$ satisfies

$$E_0(\mathbf{y}) = \boldsymbol{\mu}_0 = b'(\boldsymbol{\theta}_0) \quad \text{and} \quad Var_0(\mathbf{y}) = a(\phi)\text{diag}\{b''(\theta_{01}), ..., b''(\theta_{0n})\} = a(\phi)\mathbf{V}_0^T\mathbf{V}_0, \tag{4.70}$$

where $\mathbf{V}_0 = \text{diag}\{b''(\theta_{01})^{1/2}, ..., b''(\theta_{0n})^{1/2}\}$ is a diagonal matrix of the true scaled standard errors of y_i. The log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_i(y_i; \boldsymbol{\theta}, \phi) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right]. \tag{4.71}$$

Let $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ be an estimate of $\boldsymbol{\theta}_0$. It can be either $\hat{\boldsymbol{\theta}}^*_\alpha = \mathbf{X}\hat{\boldsymbol{\beta}}^*_\alpha$, the maximum likelihood estimate under model α, or $\hat{\boldsymbol{\theta}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda$, the penalized estimate corresponding to tuning parameter λ. Note that this $\hat{\boldsymbol{\theta}}$ depends on sample size n, but we omit the subscript n for simplification of notation.

The average Kullback-Leibler loss of an estimate can then be written as

$$L_{KL}(\hat{\boldsymbol{\beta}}) = \frac{2}{n} \sum_{i=1}^{n} \rho(\theta(\mathbf{x}_i), \mathbf{x}_i^T\hat{\boldsymbol{\beta}}) = \frac{2}{n} E_0 \left\{ \ell(\boldsymbol{\theta}_0) - \ell(\hat{\boldsymbol{\theta}}) \right\}, \tag{4.72}$$

where $\rho(\cdot, \cdot)$ is the KL discrepancy measure. Note that $\hat{\boldsymbol{\beta}}$ and hence $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are treated as nonrandom in calculating the above expectation. The KL risk is defined as $R_{KL}(\hat{\boldsymbol{\beta}}) = E[L_{KL}(\hat{\boldsymbol{\beta}})]$.

The following Lemma shows that for $\alpha \in \mathcal{C}$, $\ell(\hat{\boldsymbol{\theta}}_\alpha)$ can be asymptotically approximated by a quadratic function. From the asymptotic approximation of $\ell(\hat{\boldsymbol{\theta}}_\alpha)$, we can further show that the average KL loss can be expanded approximately as a quadratic function.

Lemma 4.4. *For $\alpha \in \mathcal{C}$, $\ell(\hat{\boldsymbol{\theta}}_\alpha)$ can be asymptotically expanded as*

$$\tilde{\ell}(\hat{\boldsymbol{\theta}}_\alpha) = \frac{1}{a(\phi)} \sum_{i=1}^{n} \left\{ y_i(\hat{\theta}_{\alpha i} - \theta_{0i}) - \left[b'(\theta_{0i})(\hat{\theta}_{\alpha i} - \theta_{0i}) + \frac{1}{2} b''(\theta_{0i})(\hat{\theta}_{\alpha i} - \theta_{0i})^2 \right] \right\}, \tag{4.73}$$

in the sense that for any bounded weights $\{w_i, i = 1, ..., n\}$ such that $0 < c_1 < w_i <$

c_2 for some constants c_1 and c_2,

$$\frac{\left|\ell(\hat{\boldsymbol{\theta}}_\alpha) - \ell(\boldsymbol{\theta}_0) - \tilde{\ell}(\hat{\boldsymbol{\theta}}_\alpha)\right|}{\sum_{i=1}^n w_i(\hat{\theta}_{\alpha i} - \theta_{0i})^2} \to 0, \tag{4.74}$$

in probability. As a special case, we let $w_i = b''(\theta_{0i})^{-1}$. Then

$$\ell(\hat{\boldsymbol{\theta}}_\alpha) = \ell(\boldsymbol{\theta}_0) + \tilde{\ell}(\hat{\boldsymbol{\theta}}_\alpha) + o_P\left(\sum_{i=1}^n b''(\theta_{0i})^{-1}(\hat{\theta}_{\alpha i} - \theta_{0i})^2\right), \tag{4.75}$$

and similarly, the average KL loss has the expansion

$$L_{KL}(\hat{\boldsymbol{\beta}}) = \frac{1}{na(\phi)}\left\{\sum_{i=1}^n b''(\theta_{0i})^{-1}(\hat{\theta}_{\alpha i} - \theta_{0i})^2 + o_P\left(\sum_{i=1}^n b''(\theta_{0i})^{-1}(\hat{\theta}_{\alpha i} - \theta_{0i})^2\right)\right\}. \tag{4.76}$$

Proof. At each θ_{0i}, we have the expansion

$$b(\hat{\theta}^*_{\alpha i}) = b(\theta_{0i}) + b'(\theta_{0i})(\hat{\theta}^*_{\alpha i} - \theta_{0i}) + \frac{1}{2}b''(\theta_{0i})(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2 + \frac{1}{6}b'''(\zeta_i)(\hat{\theta}^*_{\alpha i} - \theta_{0i})^3, \tag{4.77}$$

for some ζ_i such that $|\zeta_i - \theta_{0i}| < |\hat{\theta}^*_{\alpha i} - \theta_{0i}|$. Note that ζ_i is in a neighborhood of θ_{0i} which is an interior point of Θ. Therefore $|b'''(\zeta_i)| < K$ for some constant $K > 0$ by assumption. Hence

$$\frac{\left|\ell(\hat{\boldsymbol{\theta}}^*_\alpha) - \ell(\boldsymbol{\theta}_0) - \tilde{\ell}(\hat{\boldsymbol{\theta}}^*_\alpha)\right|}{\sum_{i=1}^n w_i(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2} = \frac{\left|\sum_{i=1}^n \frac{1}{6}b'''(\zeta_i)(\hat{\theta}^*_{\alpha i} - \theta_{0i})^3\right|}{\sum_{i=1}^n w_i(\hat{\theta}^*_{\alpha i} - \theta_{0i})^2} \le \frac{K}{6c_1}\sup_{1\le i\le n}|\hat{\theta}^*_{\alpha i} - \theta_{0i}|, \tag{4.78}$$

which goes to zero in probability when $\sup_{1\le i\le n}|\hat{\theta}^*_{\alpha i} - \theta_{0i}| \to 0$ in probability as $n \to \infty$. □

Note that although Lemma 4.4 is stated in terms of the MLE $\hat{\boldsymbol{\theta}}^*_\alpha$, it can be shown following the proof that the lemma also holds for any estimate $\hat{\boldsymbol{\theta}}$ that satisfies the

condition in 4.68. Next we introduce some notations to simplify the expressions. Denote

$$y_i^\dagger = b''(\theta_{0i})^{-1/2}(y_i - b'(\theta_{0i})) + b''(\theta_{0i})^{1/2}\theta_{0i} \quad \text{and} \quad \hat{\theta}_i^\dagger = b''(\theta_{0i})^{1/2}\hat{\theta}_i, \tag{4.79}$$

or in vector notations,

$$\mathbf{y}^\dagger = \mathbf{V}_0^{-1}(\mathbf{y} - b'(\boldsymbol{\theta}_0)) + \mathbf{V}_0\boldsymbol{\theta}_0 \quad \text{and} \quad \hat{\boldsymbol{\theta}}^\dagger = \mathbf{V}_0\hat{\boldsymbol{\theta}}. \tag{4.80}$$

When we maximizes $\ell(\boldsymbol{\theta})$ to solve for the MLE, we can drop any constant with respect to $\boldsymbol{\theta}$, and hence

$$\ell(\hat{\boldsymbol{\theta}}) = -\frac{1}{2a(\phi)} \parallel \mathbf{y}^\dagger - \hat{\boldsymbol{\theta}}^\dagger \parallel^2 (1 + o_P(1)), \tag{4.81}$$

is asymptotically a quadratic function under the conditions in Lemma 4.4. Furthermore, to simplify the KL loss, we note that $\sum_{i=1}^n b''(\theta_{0i})^{-1}(\hat{\theta}_i - \theta_{0i})^2 = \sum_{i=1}^n (\hat{\theta}_i^\dagger - \theta_{0i}^\dagger)^2$, and hence

$$L_{KL}(\hat{\boldsymbol{\beta}}) = \frac{1}{na(\phi)} \parallel \hat{\boldsymbol{\theta}}^\dagger - \boldsymbol{\theta}_0^\dagger \parallel^2 (1 + o_P(1)). \tag{4.82}$$

This means that the KL loss can be treated asymptotically as a special squared loss.

Remark 4.4 (Normal linear regression). In normal linear regression with known variance, the variance matrix $\mathbf{V}_0$ is an identity matrix, and $\boldsymbol{\theta}_0 = \boldsymbol{\mu}_0 = b'(\boldsymbol{\theta}_0)$, and hence $\mathbf{y}^\dagger$ reduces to $\mathbf{y}$, and maximizing the log-likelihood function is the same as minimizing the residual sum of squares $\parallel \mathbf{y} - \hat{\boldsymbol{\mu}} \parallel^2$. Meanwhile, the average KL loss reduces to $(n\sigma^2)^{-1} \parallel \boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}} \parallel^2$, which is linear in relationship

with the squared loss.

4.3.4 Asymptotic Loss Efficiency of GLIM

The asymptotic expansions of the likelihood function and the KL loss allow us to write the target quantities as asymptotic quadratic functions under the conditions imposed in Lemma 4.4. This makes it possible to adopt the strategies used to develop asymptotic loss efficiency for least squares to the GLIM setting. For the sake of simplicity, in the rest of this section, all the equations only consider the leading terms on both sides.

Classical generalized information criterion is carried out via best subset selection: among all the candidate models in $\mathcal{A}$, we choose the model $\hat{\alpha}$ that minimizes

$$\mathrm{GIC}^*_{\kappa_n}(\alpha) = \frac{1}{n} D(\mathbf{y}; \hat{\boldsymbol{\theta}}^*_\alpha) + \frac{1}{n} \kappa_n d_\alpha, \tag{4.83}$$

where $\hat{\boldsymbol{\theta}}^*_\alpha = \mathbf{X}\hat{\boldsymbol{\beta}}^*_\alpha$ and $D(\mathbf{y}; \hat{\boldsymbol{\theta}}^*_\alpha)$ is the scaled deviance function. When $\kappa_n = 2$, it corresponds to the classical AIC; when $\kappa_n = \log n$, it corresponds to the classical BIC.

In order to establish the asymptotic loss efficiency, we need to further restrict our focus to some candidate set $\mathcal{D} \subset \mathcal{C}$, such that

$$\sup_{\alpha \in \mathcal{D}} \sup_{1 \le i \le n} |\hat{\theta}^*_{\alpha i} - \theta_{0i}| \to 0, \tag{4.84}$$

in probability. That is, we look at the models for which the MLEs are uniformly close to the truth. The results cannot be established with only pointwise asymptotics, as shown in Lemma 4.1. Adopting the loss efficiency in Li (1987), we say a

classical criterion is asymptotically KL-loss efficient if

$$\frac{L_{KL}(\hat{\boldsymbol{\beta}}^*_{\hat{\alpha}})}{\inf_{\alpha\in\mathcal{D}} L_{KL}(\hat{\boldsymbol{\beta}}^*_{\alpha})} \to 1. \tag{4.85}$$

We first briefly discuss the KL loss of the maximum likelihood estimate of some specific candidate model α. When canonical links are used, we model the canonical parameter $\boldsymbol{\theta}_0$ with $\mathbf{X}_\alpha\boldsymbol{\beta}_\alpha$ under a candidate model α. Write $\mathbf{X}^\dagger_\alpha = \mathbf{V}_0\mathbf{X}_\alpha$; then modelling $\boldsymbol{\theta}_0$ with $\mathbf{X}_\alpha\boldsymbol{\beta}_\alpha$ is the same as modelling $\boldsymbol{\theta}^\dagger_0$ with $\mathbf{X}^\dagger_\alpha\boldsymbol{\beta}_\alpha$. From the expansion of the likelihood function, we know that the MLE of $\boldsymbol{\beta}$ under model α minimizes $\parallel \mathbf{y}^\dagger - \mathbf{X}^\dagger_\alpha\boldsymbol{\beta} \parallel^2$. Therefore in an asymptotical sense, the MLE is $\hat{\boldsymbol{\beta}}^*_\alpha = (\mathbf{X}^{\dagger T}_\alpha\mathbf{X}^\dagger_\alpha)^{-1}\mathbf{X}^{\dagger T}_\alpha\mathbf{y}^\dagger$. This expression can also be obtained from equation (4.66).

Denote by $\mathbf{H}^\dagger_\alpha = \mathbf{X}^\dagger_\alpha(\mathbf{X}^{\dagger T}_\alpha\mathbf{X}^\dagger_\alpha)^{-1}\mathbf{X}^{\dagger T}_\alpha$ the projection matrix with respect to the space spanned by $\mathbf{X}^\dagger_\alpha$. We can also write the above as $\hat{\boldsymbol{\theta}}^\dagger_\alpha = \mathbf{X}^\dagger_\alpha\hat{\boldsymbol{\beta}}^*_\alpha = \mathbf{H}^\dagger_\alpha\mathbf{y}^\dagger$. Hence, dropping the dispersion parameter, we can write the Kullback-Leibler loss of $\hat{\boldsymbol{\beta}}^*_\alpha$ as

$$\begin{aligned} a(\phi)L_{KL}(\hat{\boldsymbol{\beta}}^*_\alpha) = \frac{1}{n}\parallel \hat{\boldsymbol{\theta}}^\dagger_\alpha - \boldsymbol{\theta}^\dagger_0 \parallel^2 &= \frac{1}{n}\parallel \mathbf{H}^\dagger_\alpha\mathbf{y}^\dagger - \mathbf{H}^\dagger_\alpha\boldsymbol{\theta}^\dagger_0 \parallel^2 + \frac{1}{n}\parallel \mathbf{H}^\dagger_\alpha\boldsymbol{\theta}^\dagger_0 - \boldsymbol{\theta}^\dagger_0 \parallel^2 \\ &= \frac{\boldsymbol{\epsilon}^{\dagger T}\mathbf{H}^\dagger_\alpha\boldsymbol{\epsilon}^\dagger}{n} + \Delta^\dagger_\alpha, \end{aligned} \tag{4.86}$$

where $\boldsymbol{\epsilon}^\dagger = \mathbf{y}^\dagger - \boldsymbol{\theta}^\dagger_0 = \mathbf{V}_0^{-1}(\mathbf{y} - b'(\boldsymbol{\theta}_0))$ is the "standardized" error term whose components all have mean 0 and variance $a(\phi)$, and $\Delta^\dagger_\alpha = \frac{1}{n}\parallel \mathbf{H}^\dagger_\alpha\boldsymbol{\theta}^\dagger_0 - \boldsymbol{\theta}^\dagger_0 \parallel^2$ is the distance between the true $\boldsymbol{\theta}^\dagger_0$ and its projection on the space spanned by $\mathbf{X}^\dagger_\alpha$, which corresponds to the systematic bias of model α. Taking expectations of both sides of (4.86) and noting that $E\left(\boldsymbol{\epsilon}^{\dagger T}\mathbf{H}^\dagger_\alpha\boldsymbol{\epsilon}^\dagger\right) = tr\mathbf{H}^\dagger_\alpha E\left(\boldsymbol{\epsilon}^\dagger\boldsymbol{\epsilon}^{\dagger T}\right) = d_\alpha a(\phi)$, we also have

$$a(\phi)R_{KL}(\hat{\boldsymbol{\beta}}^*_\alpha) = \Delta^\dagger_\alpha + \frac{a(\phi)d_\alpha}{n}. \tag{4.87}$$

Next we adopt the strategies from Li (1987) to show the asymptotic loss efficiency of AIC, i.e. the GIC with $\kappa_n = 2$ in (4.83). The following technical conditions are adopted from Li (1987).

(A1) For any candidate model $\alpha \in \mathcal{A}$, the largest eigenvalue of $\frac{1}{n}\mathbf{X}_\alpha^{\dagger T}\mathbf{X}_\alpha^\dagger$ is bounded from above uniformly by some finite number.

(A2) $E|y_i|^{4q} < \infty$ for some integer q.

(A3) The risks of the maximum likelihood estimators $\hat{\boldsymbol{\beta}}_\alpha^*$ for all $\alpha \in \mathcal{C}$ satisfy

$$\sum_{\alpha \in \mathcal{C}} [nR(\hat{\boldsymbol{\beta}}_\alpha^*)]^{-q} \to 0. \tag{4.88}$$

Theorem 4.3. *Assume conditions (A1)–(A3) hold. Then $\hat{\alpha}$, the model selected by GIC_2^* from $\mathcal{D}$, is asymptotically loss efficient in the sense of (4.85).*

Proof. This proof is adapted from the proof of Theorem 2.1 in Li (1987) by noting that the components of $\boldsymbol{\epsilon}^\dagger = \mathbf{V}_0^{-1}(\mathbf{y} - b'(\boldsymbol{\theta}_0))$ are independent with mean 0 and equal variances $a(\phi)$, although they are not identically distributed. For the sake of simplicity, we denote by C a generic constant number which might differ from equation to equation. First from Lemma 4.4, we have the following approximation of $\mathrm{GIC}_2^*(\alpha)$.

$$\begin{aligned} \mathrm{GIC}_2^*(\alpha) &= -\frac{2}{n}\ell(\hat{\boldsymbol{\beta}}_\alpha^*) + \frac{2d_\alpha}{n} \approx \frac{\| \mathbf{y}^\dagger - \hat{\boldsymbol{\theta}}_\alpha^\dagger \|^2}{na(\phi)} + \frac{2d_\alpha}{n} \\ &= L_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*) + \frac{1}{a(\phi)}\left\{ \frac{\| \boldsymbol{\epsilon}^\dagger \|^2}{n} + \frac{2\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I} - \mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger}{n} + \frac{2(a(\phi)d_\alpha - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}_\alpha^\dagger\boldsymbol{\epsilon}^\dagger)}{n} \right\} \end{aligned} \tag{4.89}$$

Then consider $\mathrm{GIC}_2^*(\alpha)$, $L_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)$ and the second term in (4.89) as $C(\lambda)$, $L(\lambda)$ and $r(\lambda)$ in Lemma 4.1 respectively. We can apply Lemma 4.1 to show that Theorem

4.3 holds if we can prove that

$$\sup_{\alpha\in\mathcal{D}} \frac{\left|\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I}-\mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger\right|}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} \to 0, \tag{4.90}$$

$$\sup_{\alpha\in\mathcal{D}} \frac{\left|a(\phi)d_\alpha - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}_\alpha^\dagger\boldsymbol{\epsilon}^\dagger\right|}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} \to 0, \tag{4.91}$$

and

$$\sup_{\alpha\in\mathcal{D}} \left|\frac{L_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)}{R_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} - 1\right| \to 0. \tag{4.92}$$

To prove (4.90), by Chebyshev's inequality we have

$$P\left\{\sup_{\alpha\in\mathcal{D}} \frac{\left|\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I}-\mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger\right|}{nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)} > \delta\right\} \le \sum_{\alpha\in\mathcal{D}} \frac{E(\boldsymbol{\epsilon}^{\dagger T}(\mathbf{I}-\mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger)^{2q}}{n^{2q}\delta^{2q}R_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)^{2q}}, \tag{4.93}$$

which by Theorem 2 of Whittle (1960), is no greater than

$$C\delta^{-2q}\sum_{\alpha\in\mathcal{D}} \frac{\parallel(\mathbf{I}-\mathbf{H}_\alpha^\dagger)\boldsymbol{\theta}_0^\dagger\parallel^{2q}}{n^{2q}R_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)^{2q}} \le C\delta^{-2q}\sum_{\alpha\in\mathcal{D}}\left[nR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)\right]^{-q}, \tag{4.94}$$

which goes to zero by condition (A3). Equation (4.91) can be shown similarly by noting

$$E(a(\phi)d_\alpha - \boldsymbol{\epsilon}^{\dagger T}\mathbf{H}_\alpha^\dagger\boldsymbol{\epsilon}^\dagger)^{2q} \le Cd_\alpha^q \le \frac{CR_{KL}(\hat{\boldsymbol{\beta}}_\alpha^*)^q}{n^q}, \tag{4.95}$$

as an application of Theorem 2 of Whittle (1960) and expansion (4.87). Finally, equation (4.92) can be shown in the same manner as (4.91) in view of the expansion (4.86). □

The results for the classical AIC variable selection procedure motivate us to study the asymptotic loss efficiency of the nonconcave penalized likelihood with general penalty functions. In the range $[0, \lambda_{\max}]$, we choose the tuning parameter

$\hat{\lambda}$ that minimizes

$$\mathrm{GIC}_{\kappa_n}(\lambda) = \frac{1}{n} D(\mathbf{y}; \hat{\boldsymbol{\mu}}_\lambda) + \frac{1}{n} \kappa_n d_{\alpha_\lambda}. \tag{4.96}$$

Similarly to linear regression, we are interested in whether an estimator is asymptotically loss efficient in the sense

$$\frac{L_{KL}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}_n})}{\inf_{\lambda \in \Lambda} L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)} \to 1, \tag{4.97}$$

in probability, where $\Lambda = \{\lambda \in [0, \lambda_{\max}] : \alpha_\lambda \in \mathcal{D}\}$

In addition to the conditions (A1)-(A3), we need the following condition to regularize the penalized estimator.

(A4) Let $\flat = (\flat_1, ..., \flat_d)^T$, where $\flat_j = p'_\lambda(|\hat{\beta}_{\lambda j}|)\mathrm{sgn}(\hat{\beta}_{\lambda j})$ for all j such that $|\hat{\beta}_{\lambda j}| > 0$, and $\flat_j = 0$ otherwise, and $\hat{\beta}_{\lambda j}$ is the j-th component of the penalized estimator $\hat{\boldsymbol{\beta}}_\lambda$. In addition, let $\hat{\boldsymbol{\beta}}^*_\alpha$ be the maximum likelihood estimator of $\boldsymbol{\beta}$ obtained from model α. Then, we assume that, in probability,

$$\sup_{\lambda \in \Lambda} \frac{\| \flat \|^2}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})} \to 0. \tag{4.98}$$

Theorem 4.4. *Assume conditions (R) and (A1)-(A4) hold. Then the penalized estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$ with $\hat{\lambda}$ selected by minimizing the GIC$_2$ criterion is asymptotically loss efficient in the sense of (4.97).*

Before proving Theorem 4.4, we establish the following two lemmas. Lemma 4.5 evaluates the difference between a penalized mean estimator $\hat{\boldsymbol{\mu}}_\lambda$ and its corresponding least squares mean estimator $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}$, while Lemma 4.6 demonstrates that the losses of $\hat{\boldsymbol{\mu}}_\lambda$ and $\hat{\boldsymbol{\mu}}^*_{\alpha_\lambda}$ are asymptotically equivalent.

Lemma 4.5. *Under condition (A1),*

$$\| \hat{\boldsymbol{\theta}}_{\lambda}^{\dagger} - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}^{\dagger *} \|^2 \leq nC \| \mathfrak{b} \|^2, \tag{4.99}$$

where C is a constant number and $\mathfrak{b}$ is defined in condition (A4).

Proof. Without loss of generality, we assume that the first d_{α_λ} components of $\hat{\boldsymbol{\beta}}_{\lambda}$ and $\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*}$ are nonzero, and denote them by $\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$ and $\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)}$, respectively. Thus, $\hat{\boldsymbol{\theta}}_{\lambda}^{\dagger} = \mathbf{X}^{\dagger}\hat{\boldsymbol{\beta}}_{\lambda} = \mathbf{X}_{\alpha_\lambda}^{\dagger}\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$ and $\hat{\boldsymbol{\theta}}_{\alpha_\lambda}^{\dagger} = \mathbf{X}^{\dagger}\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*} = \mathbf{X}_{\alpha_\lambda}^{\dagger}\hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)}$. From the expansion of likelihood function in (4.81), the penalized log-likelihood function

$$Q(\boldsymbol{\beta}) \propto \frac{1}{2n} \| \mathbf{y}^{\dagger} - \mathbf{X}^{\dagger}\boldsymbol{\beta} \|^2 + a(\phi) \sum_{j=1}^{n} p_{\lambda}(|\beta_j|), \tag{4.100}$$

ignoring a constant with respect to $\boldsymbol{\beta}$. From the proofs of Theorems 1 and 2 in Fan and Li (2001), with probability tending to 1, we have that $\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$ is the solution of the following equation,

$$\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^{\dagger T}\left(\mathbf{y}^{\dagger} - \mathbf{X}_{\alpha_\lambda}^{\dagger}\boldsymbol{\beta}_{\lambda}^{(1)}\right) + a(\phi)\mathfrak{b}^{(1)} = \mathbf{0}, \tag{4.101}$$

where $\mathfrak{b}^{(1)}$ is the subvector of $\mathfrak{b}$ that corresponds to $\hat{\boldsymbol{\beta}}_{\lambda}^{(1)}$. Accordingly,

$$\hat{\boldsymbol{\beta}}_{\lambda}^{(1)} = \left(\mathbf{X}_{\alpha_\lambda}^{\dagger T}\mathbf{X}_{\alpha_\lambda}^{\dagger}\right)^{-1}\mathbf{X}_{\alpha_\lambda}^{\dagger T}\mathbf{y}^{\dagger} + \left(\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^{\dagger T}\mathbf{X}_{\alpha_\lambda}^{\dagger}\right)^{-1} a(\phi)\mathfrak{b}^{(1)} = \hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)} + \left(\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^{\dagger T}\mathbf{X}_{\alpha_\lambda}^{\dagger}\right)^{-1} a(\phi)\mathfrak{b}^{(1)}. \tag{4.102}$$

In addition, the eigenvalues of $\left(\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^{\dagger T}\mathbf{X}_{\alpha_\lambda}^{\dagger}\right)^{-1}$ are bounded under condition (A1). Hence,

$$\| \hat{\boldsymbol{\theta}}_{\lambda}^{\dagger} - \hat{\boldsymbol{\theta}}_{\alpha_\lambda}^{\dagger *} \|^2 = \| \mathbf{X}_{\alpha_\lambda}^{\dagger}(\hat{\boldsymbol{\beta}}_{\lambda}^{(1)} - \hat{\boldsymbol{\beta}}_{\alpha_\lambda}^{*(1)}) \|^2 = na(\phi)^2\mathfrak{b}^{(1)T}\left(\frac{1}{n}\mathbf{X}_{\alpha_\lambda}^{\dagger T}\mathbf{X}_{\alpha_\lambda}^{\dagger}\right)^{-1}\mathfrak{b}^{(1)} \leq nC \| \mathfrak{b} \|^2, \tag{4.103}$$

for some positive constant number C. This completes the proof.

Lemma 4.6. *If conditions (A1)—(A4) hold, then*

$$\sup_{\lambda\in\Lambda}\left|\frac{L(\hat{\boldsymbol{\beta}}_\lambda)}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}-1\right|\to 0, \tag{4.104}$$

in probability.

Proof. After algebraic simplification, we have

$$L(\hat{\boldsymbol{\beta}}_\lambda)-L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})=\frac{\parallel \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda}-\hat{\boldsymbol{\theta}}^{\dagger}_{\lambda}\parallel^2}{na(\phi)}+\frac{2(\boldsymbol{\theta}^{\dagger}_0-\hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda})^T(\hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda}-\hat{\boldsymbol{\theta}}^{\dagger}_{\lambda})}{na(\phi)}=I_1+I_2. \tag{4.105}$$

Under conditions (A1)-(A3), we know from (4.92) that

$$\sup_{\alpha\in\mathcal{C}}\left|\frac{L(\hat{\boldsymbol{\beta}}^*_\alpha)}{R(\hat{\boldsymbol{\beta}}^*_\alpha)}-1\right|\to 0. \tag{4.106}$$

This, together with condition (A4) and Lemma 4.5, implies

$$\sup_{\lambda\in\Lambda}\left|\frac{I_1}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\right|=\sup_{\lambda\in\Lambda}\left\{\frac{\parallel \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda}-\hat{\boldsymbol{\theta}}^{\dagger}_{\lambda}\parallel^2}{na(\phi)R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}-\frac{\parallel \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda}-\hat{\boldsymbol{\theta}}^{\dagger}_{\lambda}\parallel^2}{na(\phi)L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\left[\frac{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}{R(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}-1\right]\right\}\to 0, \tag{4.107}$$

Applying the Cauchy-Schwarz inequality, we next obtain

$$I_2\le\frac{2\parallel \boldsymbol{\theta}^{\dagger}_0-\hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda}\parallel\cdot\parallel \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda}-\hat{\boldsymbol{\theta}}^{\dagger}_{\lambda}\parallel}{n}=2\sqrt{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\cdot\frac{1}{\sqrt{n}}\parallel \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda}-\hat{\boldsymbol{\theta}}^{\dagger}_{\lambda}\parallel. \tag{4.108}$$

As a result, $\sup_{\lambda\in\Lambda}\left|\frac{I_2}{L(\hat{\boldsymbol{\beta}}^*_{\alpha_\lambda})}\right|\to 0$, and Lemma 4.6 follows immediately.

Proof of Theorem 4.4. From Lemma 4.1, in order to show the asymptotic loss efficiency of $\mathrm{GIC}_2(\lambda)$, it suffices to demonstrate that minimizing $\mathrm{GIC}_2(\lambda)$ is the same

as minimizing $L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)$ asymptotically. To this end, we first note that ignoring a constant,

$$\begin{aligned}
\mathrm{GIC}_2(\lambda) &\approx \frac{\| \mathbf{y}^\dagger - \mathbf{X}^\dagger \hat{\boldsymbol{\beta}}_\lambda \|^2}{na(\phi)} + \frac{2d_{\alpha_\lambda}}{n} \\
&= \frac{\| \mathbf{y}^\dagger - \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} \|^2}{na(\phi)} + \frac{\| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2}{na(\phi)} + \frac{2d_{\alpha_\lambda}}{n} \\
&= L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) + \left[L_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) - L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) \right] + \frac{1}{a(\phi)} \left\{ \frac{\| \boldsymbol{\epsilon}^\dagger \|^2}{n} + \frac{1}{n} \| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2 \right. \\
&\quad \left. + \frac{2}{n} \boldsymbol{\epsilon}^{\dagger T} (\mathbf{I} - \mathbf{H}^\dagger_{\alpha_\lambda}) \boldsymbol{\theta}^\dagger_0 + \frac{2}{n} \left(a(\phi) d_{\alpha_\lambda} - \boldsymbol{\epsilon}^{\dagger T} \mathbf{H}^\dagger_{\alpha_\lambda} \boldsymbol{\epsilon}^\dagger \right) \right\}.
\end{aligned} \tag{4.109}$$

Let

$$J_1 = L_{KL}(\hat{\boldsymbol{\beta}}_{\alpha_\lambda}) - L_{KL}(\hat{\boldsymbol{\beta}}_\lambda), \tag{4.110}$$

$$J_2 = \| \hat{\boldsymbol{\theta}}^{\dagger *}_{\alpha_\lambda} - \hat{\boldsymbol{\theta}}^{\dagger}_{\lambda} \|^2 / n, \tag{4.111}$$

$$J_3 = 2\boldsymbol{\epsilon}^{\dagger T} (\mathbf{I} - \mathbf{H}^\dagger_{\alpha_\lambda}) \boldsymbol{\mu} / n, \tag{4.112}$$

and

$$J_4 = 2(a(\phi) d_{\alpha_\lambda} - \boldsymbol{\epsilon}^{\dagger T} \mathbf{H}^\dagger_{\alpha_\lambda} \boldsymbol{\epsilon}^\dagger) / n. \tag{4.113}$$

Using Lemma 4.6 and similar arguments used in the proof of Theorem 4.3, we obtain that, in probability,

$$\sup_{\lambda \in \Lambda} \left| \frac{J_j}{L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)} \right| \to 0, \quad \text{for } j = 1, ..., 4. \tag{4.114}$$

These imply that, ignoring a constant with respect to $\hat{\boldsymbol{\beta}}_\lambda$, the difference between $\mathrm{GIC}_2(\lambda)$ and $L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)$ is negligible in comparison to $L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)$. This completes the proof. □

Chapter 5

Numerical Results

In this section, we conduct four Monte Carlo simulation studies to investigate the finite sample performance of the proposed procedures. The proposed procedures are also tested by two real data applications. All numerical results are obtained using Matlab.

5.1 Simulation Studies

5.1.1 Consistency

Our simulation study is designed to compare the performance of the AIC-type selector and BIC-type selector in terms of model sparsity and model error (ME) defined by

$$\mathrm{ME}(\hat{\boldsymbol{\beta}}) = E_{\mathbf{X}}\{\mu(\mathbf{x}^T\boldsymbol{\beta}) - \mu(\mathbf{x}^T\hat{\boldsymbol{\beta}})\}^2, \tag{5.1}$$

where the expectation is taken over a new observation of the covariate vector $\mathbf{x}$, and $\mu(\mathbf{x}) = E(y|\mathbf{x})$ which depends on $\mathbf{x}$ through $\mathbf{x}^T\boldsymbol{\beta}$ for the GLIM and the Cox

model. The model error can be approximated by

$$\mathrm{ME}(\hat{\boldsymbol{\beta}}) \approx (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E_{\mathbf{X}}[\{\mu'(\mathbf{x}^T\boldsymbol{\beta})\}^2 \mathbf{x}\mathbf{x}^T](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{5.2}$$

The term $E_{\mathbf{X}}[\{\mu'(\mathbf{x}^T\boldsymbol{\beta})\}^2 \mathbf{x}\mathbf{x}^T]$ may be of a closed form in some situations. If there is no closed form, we estimate it by Monte Carlo method. Thus, we refer to $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E_{\mathbf{X}}[\{\mu'(\mathbf{x}^T\boldsymbol{\beta})\}^2 \mathbf{x}\mathbf{x}^T](\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ as the model error of $\hat{\boldsymbol{\beta}}$ with a slight abuse of terminology. To compare the model error, we report the median of the relative model error (MRME). Here, relative model error is defined to be the ratio of a penalized estimator to that of the unpenalized MLE under the full model.

To investigate whether the penalized methods overfits and how often they select the smallest correct model, we report the average number of correct and incorrect zero components in the final estimates and how many variables they overfit. Here we will examine the performance of the SCAD procedure with the GIC selectors and compare with classical criteria. In particular, we compare SCAD-AIC and SCAD-BIC, standing for the SCAD procedure with $\kappa_n = 2$ and $\kappa_n = \log(n)$, respectively, with the best subset variable selection with (traditional AIC and BIC criteria) whose solutions are obtained via exhaustive search over all possible subsets. We also include the oracle procedure as benchmark. Here the oracle procedure means that its estimate sets the zero coefficient to be zero and estimates non-zero coefficients using the MLE of the model including only variables with the nonzero coefficients.

Example 5.1 (Logistic regression). Simulation data are randomly generated

from a logistic regression model, that is, given $\mathbf{x}$, $y \sim \text{Bernoulli}\{p(\mathbf{x}^T\boldsymbol{\beta})\}$ with

$$p(\mathbf{x}^T\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta})}, \tag{5.3}$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0)^T$ and $\mathbf{x}$ is a 12-dimensional random vector whose first 9 components are multivariate normal with covariance matrix $\Sigma = (\sigma_{ij})$ where $\sigma_{ij} = 0.5^{|i-j|}$, and whose last three components are drawn from independent Bernoulli distribution with success probability 0.5. This model is taken from Tibshirani (1996) and Fan and Li (2001) by adding more zero components. In our simulation, we consider sample size $n = 100, 200, 300$ and 400. For each case, we conduct 1000 simulations.

In logistic regression, the model error is

$$\begin{aligned} \text{ME}_{Logistic} &= E\{[p(\mathbf{x}\boldsymbol{\beta}) - p(\mathbf{x}\hat{\boldsymbol{\beta}}]\}^2) \\ &\approx (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E\{\mathbf{x}\mathbf{x}^T(p'(\mathbf{x}\boldsymbol{\beta})^2)\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \end{aligned} \tag{5.4}$$

As in our simulation study, we know the distribution of $\mathbf{X}$ and the exact value of $\boldsymbol{\beta}$, so we can calculate the expectation in the middle part in advance using a Monte Carlo method and then compare different estimators.

The simulation results are summarized in Table 5.1. We report the median of RME for SCAD-AIC/BIC and conventional AIC/BIC over 1000 simulations. Correct and incorrect zeros are labelled "C" and "I" respectively. They refer to the average number of true zeros that are correctly set as zero and non-zero coefficients that are incorrectly set as zero. We also report the proportion of underfit, exactfit and overfit with 1, 2, 3, 4 and 5 or more parameters.

From Table 5.1, we can see that all variable selection procedures reduce model error and model complexity significantly. SCAD-BIC performs the best. It is

Table 5.1. Simulation results for logistic regression model

Method	MRME (%)	Zeros C	Zeros I	Under (%)	Exact (%)	Overfitted (%) 1	2	3	4	≥ 5
n=100										
Scad-AIC	48.89	6.99	0.15	9.0	23.5	20.9	17.3	13.2	9.2	11.9
Scad-BIC	17.61	8.41	0.27	18.9	54.5	18.7	6.9	3.1	1.1	2.1
AIC	50.94	6.91	0.07	6.6	16.1	26.2	23.7	15.7	9.2	8.8
BIC	15.45	8.46	0.16	14.4	59.1	25.3	7.0	2.3	0.7	0.8
Oracle	6.20	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0
$n = 200$										
Scad-AIC	58.52	7.39	0.05	1.8	29.8	23.0	19.2	13.4	7.2	5.8
Scad-BIC	17.08	8.81	0.06	2.3	82.4	13.1	2.2	0.4	0.0	0.0
AIC	57.58	7.42	0.05	1.7	19.7	31.0	26.3	14.6	4.0	2.8
BIC	18.50	8.79	0.06	2.3	79.3	16.6	2.2	0.1	0.0	0.0
Oracle	12.74	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0
$n = 300$										
Scad-AIC	63.19	7.35	0.00	0.1	28.9	23.4	20.3	14.7	7.8	4.9
Scad-BIC	18.03	8.86	0.00	0.1	88.0	9.8	2.0	0.1	0.1	0.0
AIC	61.95	7.42	0.00	0.0	20.2	33.2	24.2	15.3	5.0	2.1
BIC	19.02	8.83	0.00	0.1	84.4	14.5	1.1	0.0	0.0	0.0
Oracle	14.18	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0
$n = 400$										
Scad-AIC	64.38	7.35	0.00	0.0	29.4	21.9	21.0	15.1	8.5	4.1
Scad-BIC	19.05	8.87	0.00	0.0	89.9	7.9	1.8	0.2	0.2	0.0
AIC	63.20	7.42	0.00	0.0	21.7	29.2	27.3	14.6	5.5	1.7
BIC	20.45	8.85	0.00	0.0	86.3	12.1	1.4	0.2	0.0	0.0
Oracle	15.88	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0

slightly better than the best subset with the BIC criterion in terms of model error and proportion of exact model fit. SCAD-AIC and the best subset variable selection with the AIC criterion perform similarly in terms of model error, but SCAD-AIC improves the AIC criterion quite a bit in terms of exact model fit.

Both SCAD-AIC and SCAD-BIC are much less computational than the AIC and BIC criteria. As sample size increases, the percentage of exact fit of SCAD-BIC and classical BIC increases towards 100%. This is consistent with Theorem 3.1(B). Furthermore, it can be seen that SCAD-BIC performs almost as well as the oracle estimator when the sample size is large. Both SCAD-AIC and classical AIC yield overfit models with nonignorable probability, which is also consistent with Theorem 3.1(A).

Example 5.2 (Poisson regression). In this example, we generate data from the Poisson regression model. Given $\mathbf{x}$, $y \sim \text{Poisson}\{\mu(\mathbf{x})\}$, where

$$\mu(\mathbf{x}) = \exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{5.5}$$

where $\boldsymbol{\beta} = (0.8, 0, 0, 1, 0, 0, 0, 0, 0, 0.6, 0, 0)^T$ and $\mathbf{x}$ is the same as that in Example 1. In this example, the sample sizes are taken to be $n = 60, 100, 200$ and 400. For each case, we replicate 1000 simulations.

To find the model error,

$$\begin{aligned} \text{ME}_{Poisson} &= E\{[\exp(\mathbf{x}\boldsymbol{\beta}) - \exp(\mathbf{x}\hat{\boldsymbol{\beta}})]^2\} \\ &\approx (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E\{\mathbf{x}\mathbf{x}^T \exp(2\mathbf{x}\boldsymbol{\beta})\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \end{aligned} \tag{5.6}$$

Again, we calculate the middle part in advance using a Monte Carlo method. Simulation results are summarized in Table 5.2, in which the notation is the same as that in Table 5.1. From Table 5.2, it can be seen that SCAD-BIC and the BIC best subset behave similarly, and outperform the SCAD-AIC and the AIC best subset. The SCAD-BIC certainly outperforms the BIC best subset in terms of computational cost. The performance of SCAD-BIC is approaching that of the

Table 5.2. Simulation results for Poisson regression model

Method	MRME (%)	Zeros C	Zeros I	Under (%)	Exact (%)	Over Fitted (%) 1	2	3	4	≥ 5
n=60										
Scad-AIC	37.88	7.13	0.02	2.3	33.8	18.2	13.1	12.7	9.0	12.3
Scad-BIC	26.95	8.28	0.06	6.3	54.2	21.7	11.0	5.8	1.9	0.7
AIC	48.69	7.66	0.02	2.4	28.8	31.1	23.7	11.0	4.1	1.3
BIC	21.79	8.64	0.04	3.7	68.9	24.7	4.7	0.5	0.1	0.0
Oracle	11.62	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0
n=100										
Scad-AIC	35.39	7.25	0.00	0.0	36.9	21.1	13.2	8.7	8.6	11.5
Scad-BIC	23.35	8.40	0.00	0.4	64.8	19.8	8.6	3.4	2.3	0.7
AIC	55.05	7.55	0.00	0.0	21.3	37.1	24.1	12.0	3.6	1.9
BIC	22.32	8.70	0.00	0.0	74.3	22.0	3.1	0.5	0.1	0.0
Oracle	13.25	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0
n=200										
Scad-AIC	33.72	7.66	0.00	0.0	47.2	18.5	12.4	9.1	5.3	7.5
Scad-BIC	22.85	8.72	0.00	0.0	79.6	14.5	4.4	0.9	0.6	0.0
AIC	58.15	7.58	0.00	0.0	21.8	36.3	25.5	11.9	3.8	0.7
BIC	22.12	8.79	0.00	0.0	81.6	16.5	1.7	0.2	0.0	0.0
Oracle	15.60	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0
n=400										
Scad-AIC	34.00	7.78	0.00	0.0	52.3	18.9	9.6	5.9	5.8	7.5
Scad-BIC	21.12	8.87	0.00	0.0	89.7	8.4	1.5	0.2	0.1	0.1
AIC	59.77	7.63	0.00	0.0	23.2	37.3	24.5	9.8	4.7	0.5
BIC	24.01	8.87	0.00	0.0	88.0	11.0	0.9	0.1	0.0	0.0
Oracle	18.64	9.00	0.00	0.0	100.0	0.0	0.0	0.0	0.0	0.0

oracle estimator as the sample size increases. These results are consistent with Theorem 3.1. They are also similar to those observed in logistic regression.

5.1.2 Efficiency

According to the theory in chapter 3, when the candidate model set contains the truth, BIC-type tuning parameter selectors are able to choose the true model with probability tending to 1, and hence largely reduce the model error. However, when the candidate model set only provides an approximation to the unknown truth, BIC-type tuning parameter selectors might be too aggressive in reducing model complexity so that the resulting fitted model might not be loss efficient. On the other hand, in the case of linear regression, the AIC selector is proven to be asymptotically loss efficient in chapter 4. The following simulation study is conducted to highlight this fundamental difference between the AIC selector and the BIC selector.

Example 5.3 (The true model is not in a set of candidate models with linear regressions). Consider a linear regression model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, where the $\mathbf{x}_i$'s are i.i.d. multivariate normal random variables with dimension 13, the correlation between x_i and x_j is $0.5^{|i-j|}$, and the ϵ_i's are i.i.d. $N(0, \sigma^2)$ with $\sigma = 4$. In addition, we partition $\mathbf{x} = (\mathbf{x}_{\text{full}}^T, \mathbf{x}_{\text{exc}}^T)^T$, where $\mathbf{x}_{\text{full}}$ contains $d = 12$ covariates of the full model and $\mathbf{x}_{\text{exc}}$ is the covariate excluded from model fittings. Accordingly, we partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\text{full}}^T, \boldsymbol{\beta}_{\text{exc}}^T)^T$, where $\boldsymbol{\beta}_{\text{full}}$ is a 12×1 vector and $\boldsymbol{\beta}_{\text{exc}}$ is a scalar. To investigate the performance of the proposed methods under various parameter structures, we let $\boldsymbol{\beta}_{\text{full}} = \boldsymbol{\beta}_0 + \gamma\boldsymbol{\delta}/\sqrt{n}$, where $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0)^T$, $\boldsymbol{\delta} = (0, 0, 1.5, 1.5, 1, 1, 0, 0, 0, 0, 0.5, 0.5)^T$, γ ranges from 0 to 10, and $\boldsymbol{\beta}_{\text{exc}} = 0.2$. Because the candidate model is the subset of the full model that contains 12 covariates in $\mathbf{x}_{\text{full}}$, the above model settings ensure that the true model is not included in the set of candidate models.

We simulate 1000 data sets with $n = 400$, 800, and 1600. To study the perfor-

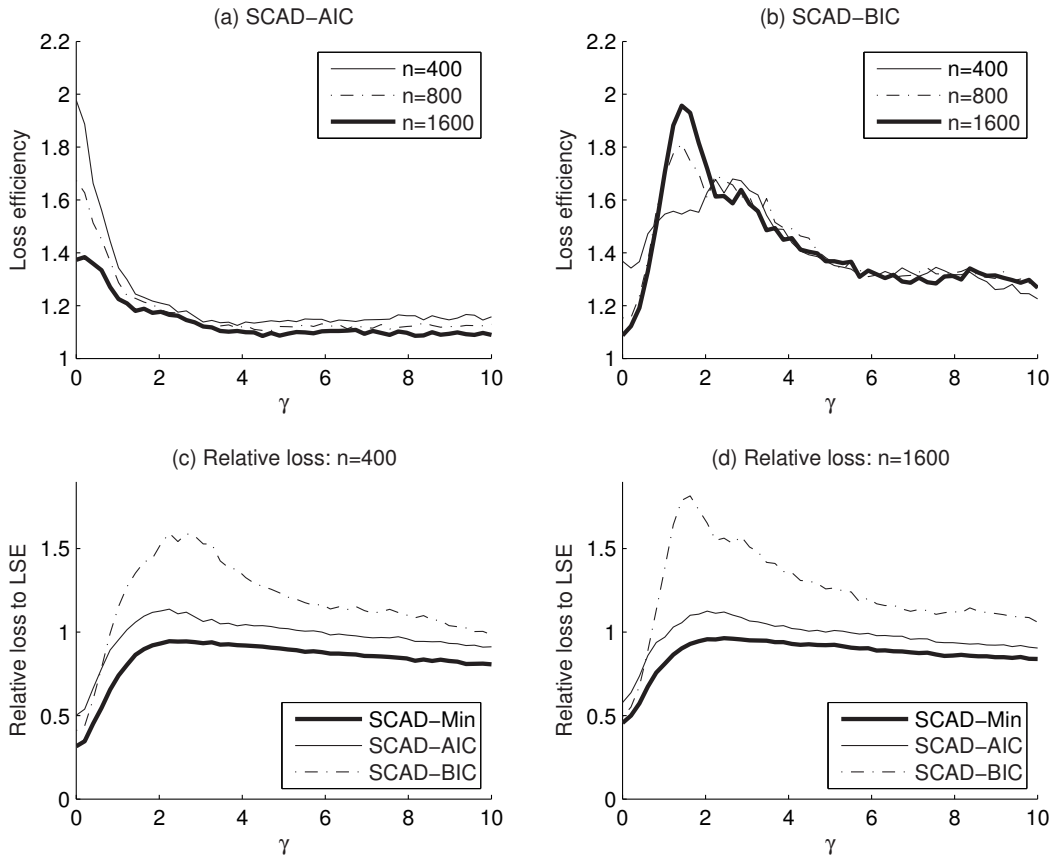

Figure 5.1. Linear regression: (a) The loss efficiency of SCAD-AIC; (b) The loss efficiency of SCAD-BIC; (c) The relative loss compared with LSE when $n = 400$; (d) The relative loss compared with LSE when $n = 1600$.

mance of selectors, we define the finite sample's loss efficiency,

$$\mathrm{LE}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}}) = \frac{L(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})}{\inf_{\lambda} L(\hat{\boldsymbol{\beta}}_{\lambda})}, \tag{5.7}$$

where $\hat{\lambda}$ is chosen by SCAD-GCV, SCAD-AIC, and SCAD-BIC. In addition, the relative loss of a penalized estimate compared with the least square estimate under the full model is defined as $RL = L(\hat{\boldsymbol{\beta}}_{\lambda})/L(\hat{\boldsymbol{\beta}}^{*}_{\bar{\alpha}})$. Because the performance of SCAD-GCV is very similar to that of SCAD-AIC, it is not reported here. Figure 5.1(a) and 5.1(b) depict the LEs of SCAD-AIC and SCAD-BIC, respectively, across

various γ. For $\lambda \in [0, \lambda_{\max}]$, we calculate the loss of $\hat{\boldsymbol{\beta}}_\lambda$ and find the optimal SCAD penalized estimate that results in the minimum loss. The relative losses of this optimal SCAD penalized estimate (SCAD-Min) and those selected by AIC and BIC-selectors are reported in Figure 5.2(c) and (d) respectively for sample sizes $n = 400$ and 1600. Figure 5.1(a) clearly indicates that the loss efficiency of SCAD-AIC converges to 1 regardless the value of γ, which corroborates the theoretical finding in Theorem 4.1. However, the loss efficiency of SCAD-BIC in Figure 5.1(b) does not show this tendency. Specifically, when γ is close to 0 (i.e., the model is nearly sparse), SCAD-BIC results in smaller loss due to its larger penalty function. As γ increases so that the full model contains the medium-sized coefficients, SCAD-BIC is likely to choose the model that is too sparse and hence increases the loss. When $\gamma \to \infty$, all coefficients are large so that we expect no variable should be excluded from the model with probability tending to 1. However, for finite sample simulation studies, although the probability is vanishing, underfitting would sometimes occur and result in large losses. As a consequence, SCAD-BIC, which employs a larger penalty, has slightly higher relative loss than SCAD-AIC when γ is large. Figures 5.2(c) and 5.2(d) show that the loss of SCAD-AIC is well controlled that the relative loss stays at a level. However, the relative loss of SCAD-BIC is higher for medium γ when sample size increases from 400 to 1600. In sum, Figure 5.1 shows that SCAD-AIC is efficient, whereas SCAD-BIC is not. Finally, we examine the efficiencies of AIC and BIC, and find that the performances of AIC and BIC are similar to those of SCAD-AIC and SCAD-BIC, respectively, and hence are omitted.

Example 5.4 (The true model is not in a set of candidate models with logistic regressions). Consider a logistic regression model $y|\mathbf{x} \sim \text{Bernoulli}\{p(\mathbf{x}^T\boldsymbol{\beta})\}$,

where

$$p(\mathbf{x}^T\boldsymbol{\beta}) = \mu(\mathbf{x}^T\boldsymbol{\beta}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1+\exp(\mathbf{x}^T\boldsymbol{\beta})}.$$

Here $\mathbf{x}_i$'s are i.i.d. multivariate normal random variables with dimension 13, the correlation between x_i and x_j is $0.5^{|i-j|}$. In addition, we partition $\mathbf{x} = (\mathbf{x}_{\text{full}}^T, \mathbf{x}_{\text{exc}}^T)^T$, where $\mathbf{x}_{\text{full}}$ contains $d = 12$ covariates of the full model and $\mathbf{x}_{\text{exc}}$ is the covariate excluded from model fittings. Accordingly, we partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\text{full}}^T, \boldsymbol{\beta}_{\text{exc}}^T)^T$, where $\boldsymbol{\beta}_{\text{full}}$ is a 12×1 vector and $\boldsymbol{\beta}_{\text{exc}}$ is a scalar. To investigate the performance of the proposed methods under various parameter structures, we let $\boldsymbol{\beta}_{\text{full}} = \boldsymbol{\beta}_0 + \gamma\boldsymbol{\delta}/\sqrt{n}$, where $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0)^T$, $\boldsymbol{\delta} = (0, 0, 1.5, 1.5, 1, 1, 0, 0, 0, 0, 0, 0.5, 0.5)^T$, γ ranges from 0 to 20, and $\boldsymbol{\beta}_{\text{exc}} = 0.2$. Because the candidate model is the subset of the full model that contains 12 covariates in $\mathbf{x}_{\text{full}}$, the above model settings ensure that the true model is not included in the set of candidate models.

We simulate 1000 data sets with $n = 400$, 800, and 1600. To study the performance of selectors, we define the finite sample's loss efficiency,

$$\text{LE}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}}) = \frac{L_{KL}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})}{\inf_\lambda L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)},$$

where $\hat{\lambda}$ is chosen by SCAD-GCV, SCAD-AIC, and SCAD-BIC. In addition, the relative loss of a penalized estimate compared with the least square estimate under the full model is defined as $RL = L_{KL}(\hat{\boldsymbol{\beta}}_\lambda)/L_{KL}(\hat{\boldsymbol{\beta}}_{\bar{\alpha}}^*)$. In our simulation, the KL loss is approximated by

$$L_{KL}(\hat{\boldsymbol{\beta}}_\lambda) \approx n^{-1}\sum_{i=1}^{n}\left[p(\mathbf{x}_{full;i}^T\hat{\boldsymbol{\beta}}_\lambda) - p(\mathbf{x}_i^T\boldsymbol{\beta})\right]^2,$$

for an estimate $\hat{\boldsymbol{\beta}}_\lambda$ from the asymptotic expansion of KL loss. Because the performance of SCAD-GCV is very similar to that of SCAD-AIC, it is not reported

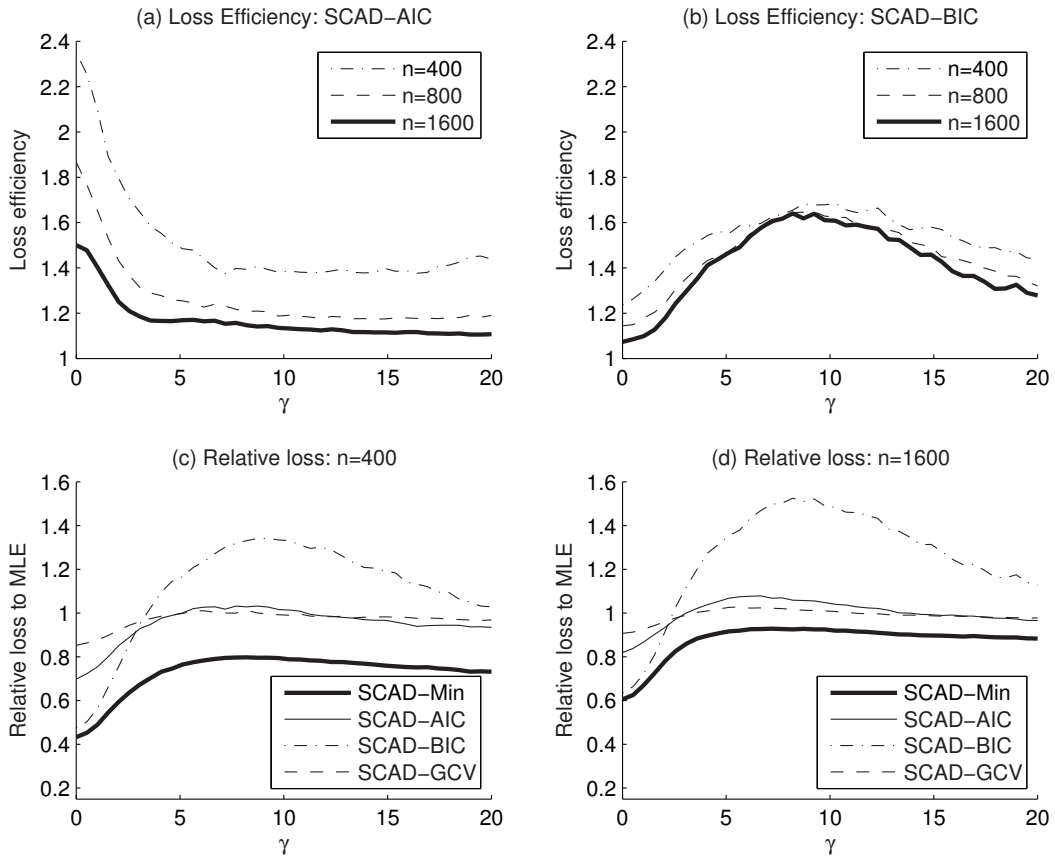

Figure 5.2. Logistic regression: (a) The loss efficiency of SCAD-AIC; (b) The loss efficiency of SCAD-BIC; (c) The relative loss compared with MLE when $n = 400$; (d) The relative loss compared with MLE when $n = 1600$.

here. Figures 5.2(a) and 5.2(b) depict the LEs of SCAD-AIC and SCAD-BIC, respectively, across various γ. For $\lambda \in [0, \lambda_{\max}]$, we calculate the loss of $\hat{\boldsymbol{\beta}}_\lambda$ and find the optimal SCAD penalized estimate that results in the minimum loss. The relative losses of this optimal SCAD penalized estimate (SCAD-Min) and those selected by the AIC and BIC selectors are reported in Figure 5.2(c) and (d) respectively for sample sizes $n = 400$ and 1600. Figure 5.2(a) clearly indicates that the loss efficiency of SCAD-AIC converges to 1 regardless of the value of γ, which corroborates the theoretical finding in Theorem 4.4. However, the loss efficiency of SCAD-BIC in Figure 5.2(b) does not show this tendency. Specifically, when γ

is close to 0 (i.e., the model is nearly sparse), SCAD-BIC results in smaller loss due to its larger penalty function. As γ increases so that the full model contains the medium-sized coefficients, SCAD-BIC is likely to choose the model that is too sparse and hence increases the loss. As a result, the LE does not decrease as n increases. Figure 5.2(c) and 5.2(d) show that the loss of SCAD-AIC is well controlled that the relative loss stays at a level. However, the relative loss of SCAD-BIC is higher for medium γ when sample size increases from 400 to 1600. In sum, Figure 5.2 shows that SCAD-AIC is efficient, whereas SCAD-BIC is not.

5.2 A Real Data Example

In this example, we analyze the mammographic mass data collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006.

Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in the last several years. These systems help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram or to perform a short term follow-up examination instead.

This data set contains 516 benign and 445 malignant instances. To predict the severity (benign/malignant), we consider the following variables: *birads*: BI-RADS assessment assigned in a double-review process by physicians, 1 (definitely benign) to 5 (highly suggestive of malignancy) (ordinal); *age*: patient's age in years (integer); *shape*: mass shape, round=1, oval=2, lobular=3, irregular=4 (nomi-

nal); *margin*: mass margin, circumscribed=1, microlobulated=2, obscured=3, ill-defined=4, spiculated=5 (nominal); *density*: mass density, high=1, iso=2, low=3, fat-containing=4 (ordinal). Missing attributes are present for some instances with the highest missing rate of 7.9% for each variable. In our analysis, we exclude cases with missing values. According to the description of the data set, the BI-RADS assessment should be between 1 and 5. However, 9 and 5 cases have a BI-RADS value of 6 and 0 respectively. We don't know if they are coded correctly and therefore excluded them. There is another case with BI-RADS assessment 55 which is suspected to be a typo of 5. We also exclude this case. Altogether there are 815 good entries for analysis. Dummy variables are created for nominal variables *shape* and *margin* with *shape* being irregular and *margin* being spiculated as baselines.

To predict the value of the binary response, $y = 0$ (benign) or $y = 1$ (malignant), on the basis of the 10 explanatory variables, we fit the data with the following logistic regression model:

$$\log \frac{p(\mathbf{x}^T\boldsymbol{\beta})}{1 - p(\mathbf{x}^T\boldsymbol{\beta})} = \beta_0 + \sum_{j=1}^{10} x_j\beta_j, \tag{5.8}$$

where $\mathbf{x} = (x_1, \cdots, x_{10})^T$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{10})^T$, and $p(\mathbf{x}^T\boldsymbol{\beta})$ is the probability of the case being classified as malignant. As a result, the tuning parameters selected by SCAD-AIC and SCAD-BIC are 0.0332 and 0.1512, respectively. SCAD-GCV results in the same model as SCAD-AIC. We also include the results from a model selected by delete-one cross validation (delete-1 CV) and 5-fold cross validation (5-fold CV). More specifically, if we obtain an estimate $\hat{\boldsymbol{\beta}}$ from some data, we define the prediction error on a new dataset $\{(\mathbf{x}_i^*, y_i^*), i = 1, ..., n^*\}$ as $\text{PE}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n^*} \left(y_i^* - p(\mathbf{x}_i^{*T}\hat{\boldsymbol{\beta}})\right)^2$. For delete-1 CV, we delete one observation from the data, fit all 2^{11} possible candidate models with the remaining data, and then calculate the

prediction error on the deleted entry. Repeat this with every observation. The final model suggested by delete-1 CV is the one with minimum average prediction error over all 815 data points. For 5-fold CV, the dataset is randomly divided into 5 parts with equal sizes. For all possible candidate models, we fit a logistic regression model with 4 parts and calculate the prediction error on the remaining validation part. We randomly generated 100 such partitions and chose the model with minimum average prediction error over 100 trials. In addition to using prediction error as a criterion for CV, we also compare average deviances over replications. We also consider the deviance instead of the prediction error as the criterion to select the best model. For this example, it turns out that both prediction error and deviance suggest an identical model for delete-1 CV and 5-fold CV.

Table 5.3. Estimates for mammographic mass data with standard deviations in parentheses

	MLE	SCAD-AIC	SCAD-BIC	5-Fold CV
β_0	-11.04(1.48)	-11.17(1.13)	-11.16(1.07)	-11.49 (1.14)
BIRADS (x_1)	2.18(0.23)	2.19(0.23)	2.25(0.23)	2.23 (0.23)
age (x_2)	0.05(0.01)	0.04(0.01)	0.04(0.01)	0.05 (0.01)
density (x_3)	-0.04(0.29)	0(-)	0(-)	0(-)
sRound (x_4)	-0.98(0.37)	-0.99(0.37)	-0.80(0.34)	-0.81 (0.35)
sOval (x_5)	-1.21(0.32)	-1.22(0.32)	-1.07(0.30)	-1.07 (0.30)
sLobular (x_6)	-0.53(0.35)	-0.54(0.34)	0(-)	0(-)
mCircum (x_7)	-1.05(0.42)	-0.98(0.32)	-1.01(0.30)	-1.07 (0.31)
mMicro (x_8)	-0.03(0.65)	0(-)	0(-)	0(-)
mObscured (x_9)	-0.48(0.39)	-0.42(0.30)	0(-)	-0.47 (0.30)
mIlldef (x_{10})	-0.09(0.33)	0(-)	0(-)	0(-)

Because the model selected by the delete-1 CV is the same as SCAD-AIC, Table 5.3 only presents the non-penalized maximum likelihood estimates from the fullmodel as well as the SCAD-AIC, SCAD-BIC, and 5-fold CV estimates, together with their standard errors. It indicates that the non-penalized maximum likelihood

approach fits five spurious variables (x_3, x_6, and x_8 to x_{10}), while SCAD-AIC and 5-fold CV include two variables (x_6 and x_9) and one variable (x_9), respectively, with insignificant effects at a level of 0.05. In contrast, all variables (x_1, x_2, x_4, x_5, and x_7) selected by SCAD-BIC are significant. Because the sample size $n = 815$ is large, these findings are consistent with Theorem 3.1 and the simulation results. In addition, the p-value of the deviance test for assessing the SCAD-BIC model against the full model is 0.41, and as a result, there is no evidence of lack of fit in the SCAD-BIC model.

Based on the model selected by SCAD-BIC, we conclude that a higher BI-RADS assessment or a greater age results in a higher chance for malignancy. In addition, the oval or round mass shape yields lower odds for malignance than does the irregular mass shape, while the odds for malignance designated by the lobular mass shape is not significantly different from that of the irregular mass shape. Moreover, the odds for malignance indicated by the microlobulated, obscured, and ill-defined mass margins are not significantly different from that of the spiculated mass margin. However, the circumscribed mass margin leads to lesser odds for malignance than that of the four other types of mass margins.

Chapter 6

Conclusion and Discussion

6.1 Conclusion Remarks

In variable selection for regression problems, classical criteria based on best subset selection are impractical for handling high-dimensional data analysis. More and more attention has been paid to the nonconcave penalized approaches which apply continuous penalties to the original object functions to perform estimation and variable selection simultaneously. The choice of the regularization parameter or tuning parameter is critical to the success of applying this methodology.

We propose the generalized information criterion (GIC) to choose regularization parameters for nonconcave penalized likelihood functions. Furthermore, we study the theoretical properties of GIC. In a very general likelihood setting, if we believe that the true model is contained in a set of candidate models, then the BIC-type selector identifies the true model with probability tending to 1, while the AIC-type selector tends to overfit with a positive probability. However, if the true model is approximated by a family of candidate models, then the AIC-type selector is asymptotically loss efficient, whereas the BIC-type selector is not, in general.

Simulation studies show that the finite sample performance of the selection criteria is in line with the theoretical properties we developed.

It is worth noting that we never intend to end the debate over which criterion being the ultimate winner. It is clear that both AIC-type and BIC-type tuning parameter selectors have their own virtues. Although the theoretical results give some guidance on the asymptotic behavior of different selectors, in practice, researchers still need to make the decision based on their own understanding of the model itself and modelling philosophy. Do they believe in a true sparse candidate model, or at least approximately? Are they willing to allow some additional complexity to better control the risk? After all, a good variable selection procedure finds the optimal fit in a set of candidate models, but it is not the remedy for a badly chosen set of candidates.

6.2 Future Work

Although much progress has been made, the properties of different tuning parameter selectors are far from clear. In this dissertation, we discuss the GIC tuning parameter selector in the classical moderate dimensional settings, i.e. the dimension of the parameter, d, is small compared with the sample size n. In many modern high dimensional problems, the number of predictors can be much larger than the available sample size, for example, in microarray data analysis. It is of great interest to study the property of different tuning parameter selectors. Intuitively, both the AIC and BIC selectors apply relatively small penalties, as both $\kappa_n = 2$ or $\log n$ is small when $d/n \to \infty$. Would both selectors be "too conservative" in the sense that too many parameters would be selected? Would RIC (risk inflation criterion, Foster & George, 1994) type selectors (e.g. $\kappa_n = \log d$) perform better than AIC

or BIC selectors? These questions are still open and worth studying in the future.

It is also worth mentioning that the efficiency of GLIM is challenging. The main difficulty is that the Kullback-Leibler loss is not quadratic except in the case of linear regression. Properly posed conditions are necessary to facilitate the Taylor expansion, which is the most powerful tool to statisticians. However, the Taylor expansion for an underfitted model is difficult to deal with, although not impossible. The conditions imposed in this dissertation for GLIM loss efficiency are not the weakest. It is of great interest whether we can weaken the assumptions or use a better formulation of the problem to better understand the loss efficiency under GLIM.

The application of nonconcave penalized variable selection methodology is not limited to the discussion in this dissertation. For example, it can be extended to partial likelihood, e.g. Cox regression models (Tibshirani, 1997; Fan & Li, 2002; Zhang & Lu, 2007), semiparametric regression models (Fan & Li, 2004; Li & Liang, 2007), and longitudinal data analysis (Fan & Li, 2002), among others. The study of tuning parameter selectors under these settings is of great interest, and requires more thorough investigation in the future.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Trans. on Automatic Control*, *19*, 716-723.

Antoniadis, A. (1997). Wavelets in Statistics: A Review (with discussion). *Journal of the Italian Statistical Association*, *6*, 97-144.

Barron, A., Birgé, L., & Massart, P. (1999). Risk Bounds for Model Selection via Penalization. *Probability Theory and Related Fields*, *113*, 301-413.

Birgé, L., & Massart, P. (2001). Gaussian Model Selection. *Journal of the European Mathematical Society*, *3*, 203-268.

Breiman, L. (1996). Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics*, *24*, 2350-2383.

Donoho, D., & Johnstone, I. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, *81*, 425-455.

Donoho, D., Johnstone, I., Kerkyacharian, G., & Picard, D. (1995). Wavelet Shrinkage; Asymptopia? *Journal of the Royal Statistical Society, Series B*, *57*, 301-369.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, *32*, 407-499.

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, *96*, 1348-1360.

Fan, J., & Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Frailty Model. *The Annals of Statistics*, *30*, 74-99.

Fan, J., & Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of the*

American Statistical Association, *99*, 710-723.

Fan, J., & Li, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.) Vol. III*, 595-622.

Fan, J., & Peng, H. (2004). Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *The Annals of Statistics*, *32*, 928-961.

Foster, D., & George, E. (1994). The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, *22*, 1947-1975.

Frank, I., & Friedman, J. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, *35*, 109-148.

Fu, W. (1998). Penalized Regressions: The Bridge Versus the LASSO. *Journal of Computational and Graphical Statistics*, *7*(3), 397-416.

Hann, E., & Quinn, B. (1979). The Determination of the Order of Autoregression. *Journal of the Royal Statistical Society, Series B*, *41*, 190-195.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models.* London: Chapman and Hall.

Hjort, N., & Pollard, D. (1993). Asymptotics for Minimisers of Convex Processes. *http://www.stat.yale.edu/˜pollard/Papers/convex.pdf.*

Hoerl, A., & Kennard, R. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*, 55-67.

Huber, P. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, *1*, 799-821.

Hurvich, C., & Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, *76*, 297-307.

Konishi, S., & Kitagawa, G. (1996). Generalised Information Criteria in Model

Selection. *Biometrika*, *83*, 875-890.

Leng, C., Lin, Y., & Wahba, G. (2006). A Note on the LASSO and Related Procedures in Model Selection. *Statistica Sinica*, *16*.

Li, K. (1987). Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, *15*, 958-975.

Li, R., Dziak, J., & Ma, Y. (2006). Nonconvex Penalized Least Squares: Characterizations, Algorithm and Application. *Manuscript*.

Li, R., & Liang, H. (2007). Nonconvex Penalized Least Squares: Characterizations, Algorithm and Application. *The Annals of Statistics.* To appear.

Mallows, C. (1973). Some Comments on C_p. *Technometrics*, *15*, 661-675.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models.* New York: Chapman & HALL/CRC.

McQuarrie, A., & Tsai, C.-L. (1998). *Regression and Time Series Model Selection.* (1st ed.). Singapore: World Scientific Publishing Co, Pte. Ltd.

Miller, A. (2002). *Subset Selection in Regression.* (Second ed.). New York: Chapman & HALL/CRC.

Nishii, R. (1984). Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression. *The Annals of Statistics*, *12*, 758-765.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *19*(2), 461-464.

Shao, J. (1997). An Asymptotic Theory for Linear Model Selection. *Statistica Sinica*, *7*, 221-264.

Shibata, R. (1980). Asymptotically Efficient Selection of the Order of the Model for Estimating Parameters of a Linear Process. *The Annals of Statistics*, *8*, 147-164.

Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrika*,

68, 45-54.

Shibata, R. (1984). Approximation Efficiency of a Selection Procedure for the Number of Regression Variables. *Biometrika*, *71*, 43-49.

Tibshirani, R. (1996). Regression Shrinkage and Selection via LASSO. *Journal of the Royal Statistical Society, Series B*, *58*, 267-288.

Tibshirani, R. (1997). The LASSO Method for Variable Selection in the Cox Model. *Statistics in Medicine*, *16*, 385-395.

Wang, H., Li, G., & Tsai, C.-L. (2007a). Regression coefficient and autoregressive order shrinkage and selection via LASSO. *Journal of the Royal Statistical Society, Series B*, *69*, 63-78.

Wang, H., Li, R., & Tsai, C.-L. (2007b). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, *94*, 553-568.

Whittle, P. (1960). Bounds For the Moments of Linear and Quadratic Forms in Independent Variables. *Theory of Probability and Its Applications*, *5*, 302-305.

Yang, Y. (2005). Can the Strengths of AIC and BIC be shared? A Conflict Between Model Identification and Regression Estimation. *Biometrika*, *92*, 937-950.

Yang, Y., & Barron, A. (1999). Information-Theoretic Determination of Minimax Rates of Convergence. *The Annals of Statistics*, *27*, 1564-1599.

Zhang, C.-H. (2007). Penalized linear unbiased selection. *Technical Report No. 2007-003. Department of Statistics, Rutgers University.*

Zhang, H. H., & Lu, W. (2007). Adaptive LASSO for Cox's Proportional Hazards Model. *Biometrika*, *94*, 691-703.

Zou, H. (2006). The Adaptive LASSO and its Oracle Properties. *Journal of the American Statistical Association*, *101*, 1418-1429.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the Degrees of Freedom of the LASSO. *The Annals of Statistics*, *35*, 2173-2192.

Vita

Yiyun Zhang

Department of Statistics, Penn State University

326 Thomas Building, University Park, PA 16802

Phone: (814) 863-1772 Email: yiyun@psu.edu

Education

Ph.D. in Statistics, The Pennsylvania State University, USA, 2009 (Expected)

M.S. in Statistics, The Pennsylvania State University, USA, 2006

B.S. in Mathematics, Fudan University, China, 2004

Awards and Honors

ENAR Distinguished Student Paper Award *2009*

Professional Experience

Reseach Assistant 07/2005 to Present

Advisor: Prof. Runze Li, Department of Statistics, PSU

Research Summer Intern 06/2006 - 08/2006 and 07/2007 - 08/2007

Advisor: Dr. Yi Tsong, Office of Biostatistics, CDER, FDA

Publications

Zhang, Y., Li, R., and Tsai, C.-L. (2008). Regularization Parameter Selections via Generalized Information Criterion. Revised for publication.

Zhang, Y. and Li, R., (2008). Iterative Conditional Maximization Algorithm for Nonconcave Penalized Likelihood. Submitted for publication.

Cole, P., Hall, S., Tan, P., **Zhang, Y.**, Crnic, K., Blair, C. and Li, R. (2008). The Development of Emotion Regulation in Early Childhood. Submitted.

Feng, Y., Li, R., Sudjianto, A. and **Zhang, Y.** (2008). Neural Network Quantile Regression with Applications to Analysis of Credit Portfolio Data. Submitted.

Conferences Presentations

2009 Regularization Parameter Selections via Generalized Information Criterion, *ENAR*, San Antonio, TX, USA.

2008 Tuning Parameter Selections for Penalized Likelihood Functions, Contributed Talk, *JSM*, Denver, CO, USA.

2007 Missing Data Analysis in Bioequivalence Trials, Topic-Contributed Talk, *JSM*, Salt Lake City, UT, USA