Applications of Penalized Likelihood Methods for Feature Selection in Statistical Modeling

by

Chen Xu

M.A., York University, Canada, 2007

B.Sc., Xi'an Jiaotong University, China, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE STUDIES
(Statistics)

The University Of British Columbia
(Vancouver)

September 2012

© Chen Xu, 2012

Abstract

Feature selection plays a pivotal role in knowledge discovery and contemporary scientific research. Traditional best subset selection or stepwise regression can be computationally expensive or unstable in the selection process, and so various penalized likelihood methods (PLMs) have received much attention in recent decades. In this dissertation, we develop approaches based on PLMs to deal with the issues of feature selection arising from several application fields.

Motivated by genomic association studies, we first address feature selection in ultra-high-dimensional situations, where the number of candidate features can be huge. Reducing the dimension of the data is essential in such situations. We propose a novel screening approach via the sparsity-restricted maximum likelihood estimator that removes most of the irrelevant features before the formal selection. The model after screening serves as an excellent starting point for the use of PLMs. We establish the screening and selection consistency of the proposed method and develop efficient algorithms for its implementation.

We next turn our attention to the analysis of complex survey data, where the identification of influential factors for certain behavioral, social, and economic indices forms a variable selection problem. When data are collected though survey sampling from a finite population, they have an intrinsic dependence structure and may provide a biased representation of the target population. To avoid distorted conclusions, survey weights are usually adopted in these analyses. We use a pseudo-likelihood to account for the survey weights and propose a penalized pseudo-likelihood method for the variable selection of survey data. The consistency of the proposed approach is established for the joint randomization framework.

Lastly, we address order selection for finite mixture models, which provides a flexible tool for modeling data from a heterogeneous population. PLMs are attractive for such problems. However, this application requires maximizations over nonsmooth and nonconcave objective functions, which are computationally challenging. We transform the original multivariate objective function into a sum of univariate functions and design an iterative thresholding-based algorithm to efficiently solve the sparse maximization without ad hoc steps. We establish the convergence of the new algorithm and illustrate its efficiency through both simulations and real-data examples.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First, I would like to express my sincere gratitude to my research supervisor Dr. Jiahua Chen for his invaluable suggestions, guidance, patience, and continuing support. I have learnt much from him, particularly regarding academic research and scientific writing, which will surely benefit my future career.

Secondly, I would like to extend my appreciation to my supervisory committee, Dr. Ruben Zamar and Dr. Lang Wu, for their suggestions and encouragement throughout my Ph.D. Also, I will never forget the enjoyable pingpong time we had together.

Finally, I would like to thank Peggy, Viena, Elaine, and Andrea for their warm and thoughtful assistance with administrative issues. Peggy and Andrea in particular helped me very much. Thank you for all the conversations and the personal advice.

Chen Xu
September 2012

Chapter 1

Introduction

1.1 Overview

Technological innovations have had a profound impact on the process of knowledge discovery. It is now feasible to collect data of unprecedented size and complexity in diverse areas of scientific research. For example, in computational genomics, geneticists may measure hundreds or thousands of gene expressions to identify the few that are associated with major diseases (Keller et al. [2009]). In market research, long-term economic data are often used to discover subgroups of customers with different needs and consumption behaviors (Dickson and Ginter [1987]). In internet applications, huge numbers of uniform resource locators (URLs) are often analyzed to learn the rules for detecting web links with pop-up advertisements (Kushmerich [1999]). Other examples occur in bioinformatics, geology, neurology, health science, economics, and finance. Although the objectives differ in various disciplines, explaining the variation in the variable of interest is a common need in their research.

Statistical modeling is one of the most powerful and widely used mathematical tools for data analysis. It aims to provide insightful summaries of the information available and to formulate learning rules based on the observed data. A statistical model mimics the generation of data through a constructed stochastic procedure and attempts to explain the variable variation via a mathematical formulation. For instance, a regression model can be used to describe the relationship between a disease status and gene expressions, and a finite mixture model helps to explain the heterogeneity of consumer behavior. Other classical applications of statistical modeling include graphical models in independence structure learning, proportional hazard models in survival analysis, and autoregressive models in time series forecasting.

When no prior knowledge is available, researchers may consider many potential variables at the initial stage of the modeling. Examples include a regression model with a huge number of covariates, an autoregressive model with a high order, and a finite mixture model with a large number of mixing components. A sophisticated model with many variables provides a better descriptive value for the data structure, but it often leads to low predictive accuracy and poor model interpretability. Hence, we wish to identify important features in massive data and to produce a parsimonious model.

The issue of feature (model) selection has been studied for decades. Traditional selection procedures, such as best subset selection and stepwise regression, are typically designed for conventional settings, where the observations are assumed to be independent and the number of candidate features is relatively small. However, contemporary scientific research often encounters datasets with high dimensionality (a huge number of variables) and/or a complex nonindependent structure. These attributes challenge traditional methods in terms of both theoretical optimality and practical feasibility. For instance, when best subset selection is used in a regression with thousands of covariates, it can be computationally infeasible and unstable in the selection process. Therefore, we need innovative selection approaches that are suitable for the new environment.

The penalized likelihood method (PLM) has been demonstrated to be an attractive technique for feature selection. Examples include the least absolute shrinkage and selection operator (LASSO) proposed by Tibshirani [1996] and the smoothly clipped absolute deviation (SCAD) introduced by Fan and Li [2001]. These approaches exclude variables from the model by estimating their coefficients to be zero and shrink the other coefficients accordingly. Compared with traditional methods, the PLM has a lower computational cost and provides more stable selection results. The PLM has received a great deal of attention and covered a wide range of statistical models. In this dissertation, we address new research problems regarding feature selection arising from different applications of the PLM.

In Section 1.2, we outline a few feature selection problems from computational genomics, chronic disease studies, and market research as motivating examples for our research. In Section 1.3, we give a brief review of traditional approaches and discuss their limitations in the context of modern data analysis. We then, in Section 1.4, introduce the PLM framework and discuss its properties for variable selection. We discuss both the theoretical and computational aspects, providing the necessary background for the detailed discussion in later chapters. In Section 1.5 we state the aim of our research and outline the contributions of this dissertation.

1.2 Motivating examples

Feature selection is fundamental to contemporary scientific research. We begin with a few motivating examples from various scientific fields to illustrate the challenges for feature selection in the context of modern data analysis.

1.2.1 Computational genomics

In contemporary biomedical studies, a common goal is to find the genetic explanations (e.g., genes) that are responsible for observable traits such as blood pressure, height, or susceptibility to disease. Understanding the genetic associations of diseases helps medical researchers to further investigate the diseases and to develop the corresponding treatment methods. In contrast to simple observable traits such as gender or blood type, some complex diseases (e.g., leukemia or diabetes) are believed to be the result of many genetic and environmental factors. Since the influential genes may spread over the whole DNA sequence (the genome), studies of complex diseases normally require a full genome scan on all possible genes. Geneticists often measure hundreds or thousands of genes for a relatively small number of participants.

In an ongoing project conducted by the Faculty of Dentistry, University of British Columbia, about 600 microRNA (miRNA) expressions in serum samples were measured from two groups of participants. One group consisted of 30 oral-cancer patients and the other group consisted of 26 individuals without cancer. The question is whether these miRNA readings can be used to distinguish the cancer patients from the others. If the method is successful, the genetic information might be further used to predict whether an oral-cancer patient will progress from a minor tumor to a serious one. Using all 600 miRNAs for the classification leads to a poor predictive value because of the high level of noise. Consequently, it is important to select those that make the greatest contribution to identifying oral-cancer patients. To this end, a traditional two-sample t-test procedure can be carried out to detect the miRNAs that have significant expression differences between the two groups. However, for the simultaneous testing of 600 genetic readings, classical methods to control the probability of false discoveries are no longer relevant. Advanced adjustments are often needed to control the false-discovery rates. Another common strategy is to build a logistic regression of the tumor type on the miRNA readings; we can then identify the relevant miRNAs by selecting the most important regression covariates. However, the number of covariates p is 600, and the number of participants n is just 56. This large-p-small-n situation places this problem outside the domain of classical model selection methods (see Section 1.3.2). We need innovative methods to deal with the high dimensionality.

1.2.2 Chronic disease studies

Chronic diseases are the leading cause of death in North America. Chronic conditions such as kidney disease, cardiovascular disease, anemia, and dementia result in long-term or permanent disability for millions of people, with serious quality-of-life consequences for them and their families. The Public Health Agency of Canada explores the experiences of Canadians with chronic health conditions by conducting the Survey on Living with Chronic Diseases in Canada (SLCDC) on the targeted population. One of the main objectives of SLCDC is to identify health behaviors that influence disease outcomes, so that the government can better plan and provide health services for people with chronic diseases.

Regression models are conventionally used in the analysis of SLCDC data; the goal is to detect the influential factors through a variable selection procedure. However, when variable selection is applied to survey data, many potential complications arise. First, the data collected through survey sampling are usually obtained from a finite population without replacement, and hence they have an intrinsic dependence structure (i.e., non-i.i.d.). Second, in complex survey designs such as the SLCDC, the inclusion probabilities of sampling units often vary across the target population. Consequently, the correlation between the response and the covariates seen in the sample can be different from that of the population. Ignoring the survey design in the selection process may result in biased results. These special features of survey sampling reduce the effectiveness of traditional selection methods that are developed for i.i.d. sample situations. We need new methods that take into account the special features.

1.2.3 Market segmentation

In 2008 a BC marketing company collected data consisting of the annual dining-out expense for 1679 households randomly sampled in BC, Canada. The goal of their study was to identify potential subgroups of customers with different needs and consumption behaviors. A customer grouping strategy (i.e., market segmentation) is important for restaurant managers, because providing food services that are suitable for different customers often leads to higher profits.

To identify the proper segmentation, cluster analysis is often used: customers with similar characteristics are assigned to a homogeneous group. Finite mixture models are among the most powerful and widely used clustering tools. They divide the overall heterogeneous population into a mixture of several subpopulations (components), where each subpopulation represents a single cluster. A mixture model with an excessive number of components (a high order) usually overfits the data and has poor interpretive value. Determining the appropriate number of components (order selection) is crucial in applications such as market segmentation. Because of the nonregularity of finite mixture models, the classical selection criteria are no longer optimal. We need new selection methods that take into account the special features of finite mixture models.

1.3 Traditional feature selection methods

The issue of feature selection has received much attention. Classical selection approaches typically consist of two parts: a selection criterion for the comparison of models with different sets of variables (features) and an associated implementation.

Traditional selection methods are typically designed for low-dimensional and independent settings. However, an understanding of their principles, working schemes, and limitations is of vital importance for the development of new methodologies. In this section, we briefly review two classical selection criteria, the Akaike information criterion (AIC; Akaike [1973]) and the Bayesian information criterion (BIC; Schwarz [1978]), and discuss their corresponding implementations.

1.3.1 Selection criteria: AIC and BIC

Suppose the data $\{d_i = (y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ are collected independently, where y_i is the ith observation of the response variable and $\boldsymbol{x}_i = \{x_{i1}, \ldots, x_{ip}\}^T$ is the associated p-dimensional covariate vector. In a typical regression context, $(y_i, \boldsymbol{x}_i)$ is assumed to be a random sample from the population $D = (Y, \boldsymbol{X})$, where the conditional mean of Y depends on a linear form $\boldsymbol{X}\boldsymbol{\beta}$ with coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. In applications, often the effects of many covariates are unimportant, so we may assume that the corresponding coefficients are zero. Feature selection aims to identify all the covariates with nonzero coefficients. This procedure is also referred to as variable selection.

Under a more general framework, suppose the data $\boldsymbol{d} = (d_1, \ldots, d_n)$ are generated from an unspecified density function $f(d; \boldsymbol{\theta}^*)$ with a q-dimensional parameter vector $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_q^*)^T$. Usually, we are uncertain about the true density $f(d; \boldsymbol{\theta}^*)$ and instead assume a larger family of models $f(d; \boldsymbol{\theta})$, in which $\boldsymbol{\theta}^*$ is a nonvanishing subvector of the p-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$. The goal of feature selection is to estimate the dimension of the true model by comparing candidate models with different dimensions. In the literature this is often referred to as model selection.

For convenience of presentation, we do not distinguish the terms feature selection, variable selection, and model selection in the rest of this chapter. We use a unified notation s to indicate a subset of $\{1, \ldots, p\}$, which represents a candidate model with parameters $\boldsymbol{\theta}_s = \{\theta_j : j \in s\}$, and denote by s^* the true model with θ^*. Also, we use $\tau(s)$ to indicate the dimension of $\boldsymbol{\theta}_s$. Clearly, $\tau(s^*) = q$.

Under the model settings described above, the log-likelihood function of $\boldsymbol{\theta}$ is

$$l(\boldsymbol{\theta}; \boldsymbol{d}) = \sum_{i=1}^{n} \log f(d_i; \boldsymbol{\theta}). \tag{1.1}$$

AIC and BIC are members of a more general family of penalized model-fit statistics (referred to as "GIC"), applicable to a wide range of statistical models fitted by the maximum likelihood method, which takes the form

$$\text{GIC}(s) = -2l(\hat{\boldsymbol{\theta}}_s; \boldsymbol{d}) + c\tau(s), \tag{1.2}$$

where $\hat{\boldsymbol{\theta}}_s$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}(s)$ based on s and c is a positive constant that differs from one model selection criterion to another. The model with the smallest GIC is selected. The magnitude of the GIC is not generally interpretable, but differences between GIC values for different models are of interest. It can be seen that the GIC is a decreasing function of the maximized log-likelihood and an increasing function of the number of variables included in the model. Hence, a lower GIC implies either a simpler model (fewer variables), a better fit (higher maximized likelihood), or both. A model that balances complexity and goodness of fit is preferred.

With $c = 2$ and $c = \log n$, we obtain AIC and BIC respectively:

$$\begin{aligned} \text{AIC}(s) &= -2l(\hat{\boldsymbol{\theta}}_s; \boldsymbol{d}) + 2\tau(s) \qquad (1.3) \\ \text{BIC}(s) &= -2l(\hat{\boldsymbol{\theta}}_s; \boldsymbol{d}) + \log n \cdot \tau(s). \qquad (1.4) \end{aligned}$$

Note that the penalty for BIC grows with the sample size, while that for AIC remains constant. When $n \geq 8$, the penalty for BIC is larger than that for AIC, and therefore BIC tends to select models with fewer variables. Although AIC and BIC have similar forms as shown in (1.2), they are based on different statistical considerations.

AIC

Suppose as before that we have a set of candidate models S under consideration, with each $s \in S$ having parameters $\boldsymbol{\theta}_s$ to be estimated from the data. As defined in (1.1), the model s infers the probability distribution $f_s(\boldsymbol{d}) = \prod_{i=1}^n f(d_i; \boldsymbol{\theta}_s)$ for the observations $\boldsymbol{d}$. This serves as an approximation to the true distribution $f_*(\boldsymbol{d}) = \prod_{i=1}^n f(d_i; \boldsymbol{\theta}^*)$. From this point of view, the "best" model is the one that provides the most accurate approximation of $f_*(\boldsymbol{d})$.

The Kullback-Leibler information is a measure of the distance between two distributions, representing the information "lost" when the second distribution is used to approximate the first. The AIC approach applies the Kullback-Leibler information to the difference between $f_*(\boldsymbol{d})$ and $f_s(\boldsymbol{d})$:

$$\begin{aligned}
\mathcal{I}(f_*, f_s) &= \int f_*(\boldsymbol{d}) \log \frac{f_*(\boldsymbol{d})}{f_s(\boldsymbol{d})} d(\boldsymbol{d}) \\
&= \int f_*(\boldsymbol{d}) \log f_*(\mathbf{y}) d(\boldsymbol{d}) - \int f_*(\boldsymbol{d}) \log f_s(\boldsymbol{d}) d(\boldsymbol{d}) \\
&= E[\log f_*(\boldsymbol{d})] - E[\log f_s(\boldsymbol{d})]. \qquad (1.5)
\end{aligned}$$

The best model is then the model s that minimizes the Kullback-Leibler loss (1.5). Note that $E[\log f_*(\boldsymbol{d})] = \phi$ is a constant that does not depend on the model and is therefore irrelevant to the model comparison. The second term $E[\log f_s(\boldsymbol{d})]$ can be approximated by the maximized log-likelihood $l_n(\boldsymbol{d}; \hat{\boldsymbol{\theta}}_s)$ with an asymptotic bias approximately equal to the number of variables in model s. In other words, we have

$$\mathcal{I}(f_*, f_s) \approx \phi - l_n(\hat{\boldsymbol{\theta}}_s; \boldsymbol{d}) + \tau(s) = \phi + \frac{1}{2}\text{AIC}(s),$$

which implies that a comparison of the Kullback-Leibler losses (1.5) is approximately equivalent to a comparison of the corresponding AIC values.

It can be seen that the AIC approach aims to minimize the Kullback-Leibler deviation between the true distribution $f_*(\boldsymbol{d})$ and the distribution $f_s(\boldsymbol{d})$ under a particular candidate model s. Therefore, in some sense, the model selected by AIC achieves the "best" possible approximation to $f_*(\boldsymbol{d})$ over the candidate models, even when the true model s^* is not included in the model space S. However, it is well known that AIC is not a consistent selection criterion, since it does not correctly select the true model s^* with probability approaching 1 in large samples when s^* is included in model space S. For further discussion of the consistency and efficiency of AIC, see Shibata [1983], Shao [1997], and Yang [2005].

BIC

BIC has its origin in the Bayesian framework; it compares the degree of support in the data for two models. Suppose as before that we have a set of candidate models S for the observations $\boldsymbol{d}$, and each model $s \in S$ has a parameter vector $\boldsymbol{\theta}_s$ with $\tau(s)$ elements to be estimated. Under model s, the density of $\boldsymbol{d}$ with a given value of $\boldsymbol{\theta}_s$ is $f_s(\boldsymbol{d}; \boldsymbol{\theta}_s) = \prod_{i=1}^{n} f(d_i; \boldsymbol{\theta}_s)$. Assume that, conditioning on s, the prior density of $\boldsymbol{\theta}_s$ is $\pi(\boldsymbol{\theta}_s)$. Then, the marginal density of $\boldsymbol{d}$ under model s is

$$P(\boldsymbol{d}|s) = \int f_s(\boldsymbol{d}; \boldsymbol{\theta}_s)\pi(\boldsymbol{\theta}_s) d\boldsymbol{\theta}_s,$$

and the posterior probability of s given $\boldsymbol{d}$ is

$$P(s|\boldsymbol{d}) = \frac{P(\boldsymbol{d}|s)P(s)}{\sum_{s \in S} P(s)P(\boldsymbol{d}|s)},$$

where $P(s)$ denotes the prior probability of model s. The model with the highest posterior probability $P(s|\boldsymbol{d})$ is then considered to receive the most support from the data. Since $\sum_{s \in S} P(s)P(\boldsymbol{d}|s)$ is a constant for any choice of the model, choosing a model with the highest $P(s|\boldsymbol{d})$ is equivalent to choosing a model that maximizes $P(\boldsymbol{d}|s)P(s)$. Under some regularity conditions on the density $f_s(\boldsymbol{d})$, $-2\log\{P(\boldsymbol{d}|s)\}$ has a Laplace approximation given by BIC (1.4) up to an additive constant. Therefore, with uniform prior settings on s (i.e., $P(s)$ is constant for $s \in S$), the BIC approach is approximately equivalent to comparing the posterior probabilities $P(s|Y)$.

Under some conventional assumptions such as the independence structure of $\boldsymbol{d}$ and a fixed data dimension p, Rao and Wu [1989] established the consistency of BIC by showing that it is asymptotically minimized by the true model s^*. However, as discussed in Section 1.2, scientific studies often encounter data with high dimensionality and a dependence structure; these features impact the efficiency and theoretical optimality of BIC.

To show that BIC can be unsatisfactory, let us consider the genomic example of Section 1.2, where the number of candidate covariates p is 600. Specifically, let the model space be partitioned into subclasses according to the number of covariates in a model. Then, under the constant prior setting, the probability assigned to a subclass is proportional to its size. In particular, the subclass of models containing a single covariate, S_1, has size 600, while the subclass of models containing two covariates, S_2, has size $600 \times 599/2$. Thus, the prior probability assigned to S_2 is $599/2$ times that assigned to S_1. It can also be seen that the prior assigned to S_j increases almost exponentially as j increases to $p/2 = 300$. Consequently, the subclasses of large models receive much greater priors than those with small models. This encourages the selection of large models in the large-p situation, which is strongly against the principle of parsimony. Therefore, BIC is often unsatisfactory for high-dimensional data analysis.

1.3.2 Selection procedures

Best subset selection

In best subset selection, a selection criterion such as AIC or BIC is evaluated for each candidate model, and the model with the best "score" is selected. Best subset selection is effective when there are only a few candidate models; it is impractical for a large number of variables. In fact, even for datasets with a moderate number of variables, the total number of candidate models can be too large to be manageable. For example, suppose the number of covariates p in a regression model is 30. This implies that there are $2^{30} \approx$ a billion possible candidate models to explore, which is obviously a computationally expensive task.

Stepwise selection

Stepwise selection is another commonly used procedure for feature selection; it typically includes forward selection and backward elimination. In the regression context, the forward selection starts with a null model with no covariates, and then adds to the model the covariate that is most correlated with the current residual. The procedure builds a sequence of models by successively including one covariate at a time up to a prespecified number of steps or until all the covariates are included. In contrast, the backward elimination begins with the full model including all covariates and then forms a sequence of models by deleting one covariate at each step. The models in the sequence are then assessed according to some selection criterion. Compared with best subset selection, stepwise selection avoids exhaustive comparisons of all the candidate models and therefore has a lower computational cost. When $p = 30$, there are only about $\frac{1}{2}p^2 = 450$ models to be considered. However, stepwise methods have been found to be unstable in the selection process: a small change in the data could cause a very different selection result (Breiman [1995]). This is partially because once a covariate has been added to (removed from) the model at any step in the stepwise selection procedure, it is never removed from (returned to) the final model. Therefore, for complex data, the stepwise methods can easily lead to a locally optimal model. Consequently, selection results based on stepwise methods can be unreliable in practice.

1.4 Penalized likelihood methods

Various penalized likelihood methods (PLMs) have been developed for the purpose of feature selection. These methods include the least absolute shrinkage and selection operator (LASSO; Tibshirani [1996]), bridge regression (Fu [1998]), the elastic net (Zou and Hastie [2005]), and the smoothly clipped absolute deviation (SCAD; Fan and Li [2001]). The shrinkage idea of PLM has been demonstrated to cope well with many of the challenging features of contemporary data analysis. Compared with traditional methods, the PLM possesses two major advantages. First, its selection procedure is continuous, and hence it provides more robust selection results; second, it is computationally efficient, which is crucial for applications with high-throughput data. In this section, we introduce the PLM framework and discuss the theoretical and computational issues.

1.4.1 The penalized likelihood and penalty functions

Given the model settings in Section 1.3.1, the penalized likelihood is defined as

$$Q(\boldsymbol{\theta}) = l(\boldsymbol{d}; \boldsymbol{\theta}) - n \sum_{j=1}^{p} \phi_\lambda(|\theta_j|), \tag{1.6}$$

where $\phi_\lambda(.)$ is a penalty function indexed by a tuning parameter λ controlling the amount of regularization in $\boldsymbol{\theta}$. Maximizing the penalized likelihood (1.6) results in a maximum penalized likelihood estimator (MPLE) $\hat{\boldsymbol{\theta}}_\lambda$ for $\boldsymbol{\theta}$.

The form of $\phi_\lambda(.)$ determines the general behavior of $\hat{\boldsymbol{\theta}}_\lambda$. The L_0 regularization, i.e., $\phi_\lambda(|\theta|) = \lambda I(|\theta \neq 0|)$, penalizes the number of variables included in a candidate model and produces a sparse estimation for $\boldsymbol{\theta}$. For models of the same size (number of variables), the one that maximizes the unpenalized likelihood is preferred. However, the L_0 penalty is not continuous, and the maximization of the corresponding penalized likelihood coincides with the GIC-based best subset selection procedure, which requires exhaustive search and is therefore computationally demanding.

With the L_1 penalty $\phi_\lambda(|\theta|) = \lambda|\theta|$, we obtain the LASSO, which continuously shrinks the model parameters toward zero as the tuning parameter λ increases. Because of the singularity of the L_1 penalty at the origin, some parameters can be shrunk to exact zero when λ is sufficiently large. Thus, LASSO qualifies as a variable selection operator. Because of the continuity of the shrinkage process, LASSO often leads to a more stable selection result than best subset selection does. In addition, the continuous shrinkage often improves the predictive ability of the model because of the bias-variance trade-off (Tibshirani [1996]).

It is well known that the L_2 penalty $\phi_\lambda(|\theta|) = \lambda|\theta|^2$ results in a ridge regression, which continuously shrinks the model parameters toward zero but does not set them exactly to zero, and hence is not suitable for variable selection purposes. We also observe that the L_γ penalty $\phi_\lambda(|\theta|) = \lambda|\theta|^\gamma$ with $0 < \gamma < 2$ leads to a bridge regression (Frank and Friedman [1993], Fu [1998]). The bridge shrinkage is continuous only when $\gamma \geq 1$, while a sparse solution can be obtained only when $\gamma \leq 1$.

Fan and Li [2001] advocate penalty functions that give estimators (MPLEs) with the following three desirable properties:

- **Sparsity**: The estimator should automatically set small estimated model parameters to zero to reduce the model complexity.

- **Unbiasedness**: The estimator should have low bias, especially when the true value of the model parameter is large.

- **Continuity**: The estimator should be continuous to avoid instability in the model selection and prediction.

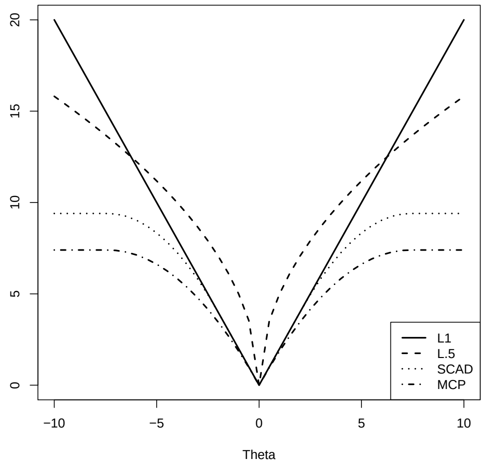

Figure 1.1: Some commonly used penalty functions, with $\lambda = 2$ for the l_1 penalty; $\lambda = 5$ for the $l_{0.5}$ penalty; ($\lambda = 2$, $a = 3.7$) for the SCAD and MCP penalties.

Of the three requirements, sparsity is the most crucial for feature extraction and variable selection, while continuity improves the robustness of the selection process. Unbiasedness is mainly required for the prediction issue, where the accuracy of the parameter estimates is the major concern. In general for a penalty function, singularity at the origin is required to generate a sparse MPLE, while concavity is needed to reduce the estimation bias.

It is known that the L_γ penalty with $\gamma > 1$ does not have sparsity, L_1 does not have unbiasedness, and L_γ with $\gamma < 1$ does not have continuity. Therefore, none of the L_γ penalties possesses all three properties. Fan and Li [2001] suggested the smoothly clipped absolute deviation (SCAD) penalty, which is defined through the following derivative

$$\phi'_\lambda(|\theta|) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\} \tag{1.7}$$

for some $a > 2$ and $\phi_\lambda(0) = 0$. As shown in Figure 1.1, the SCAD penalty takes off from the origin as the L_1 penalty and then gradually levels off as the model parameter increases. These features ensure that SCAD maintains the sparsity and continuity of l_1, and it is unbiased because it does not apply excessive shrinkage to the model parameters. Another penalty in a similar spirit is the minimax concave penalty (MCP) proposed by Zhang [2010], the derivative of which is given by

$$\phi'_\lambda(|\theta|) = \frac{(a\lambda - \theta)_+}{a}. \tag{1.8}$$

The MCP differs from SCAD for small values of θ with a strictly decreasing derivative from the origin, which further discourages the over-regularization of the model parameters. In the literature, many other penalty functions have been proposed. To account for grouping effects, Zou and Hastie [2005] suggested a linear combination of the L_1 and L_2 penalties and called the associated PLM the elastic net. To improve the estimation accuracy, Zou [2006] investigated the use of a weighted L_1 penalty and proposed the adaptive LASSO. Figure 1.1 depicts some of these commonly used penalty functions.

The shrinkage idea of PLM can also be realized through a full Bayesian analysis. ¿From a Bayesian point of view, the penalized likelihood estimator $\hat{\boldsymbol{\theta}}_\lambda$ can be interpreted as the posterior mode estimate when the model parameters $\boldsymbol{\theta}$ have correspondingly informative priors. In particular, it is well known that the L_1 penalty in LASSO corresponds to an i.i.d. Laplace (i.e., double-exponential) prior of the coefficients (Park and Casella [2008]). Compared with the PLM, Bayesian shrinkage provides more convenient interval estimates of model parameters (i.e., from the estimated posterior), but usually at a cost in terms of computational efficiency. In this dissertation, we focus on the PLM and leave the potential issues of Bayesian shrinkage methods to future research.

1.4.2 Asymptotic properties of PLM

The sampling properties of PLM have been extensively studied. The consistency of the MPLE $\hat{\boldsymbol{\theta}}_\lambda$ for feature selection and parameter estimation is of particular interest. These two modes of consistency are defined by

- **Estimation Consistency**: $\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\|_2 \to_p 0, \quad \text{as } n \to \infty,$
- **Selection Consistency**: $P(\{j : \hat{\theta}_{j\lambda} \neq 0\} = s^*) \to 1, \quad \text{as } n \to \infty,$

where $\|.\|_2$ denotes the Euclidean norm. Estimation consistency implies that as the sample size increases the MPLE $\hat{\boldsymbol{\theta}}_\lambda$ approaches the true value $\boldsymbol{\theta}^*$ with probability tending to 1, which is desirable for the parameter estimation. On the other hand, selection consistency means that, with probability tending to 1, $\hat{\boldsymbol{\theta}}_\lambda$ eliminates unimportant features by estimating their coefficients at zero, which is essential for a good feature selection method. An estimator that is consistent in terms of the parameter estimation does not necessarily consistently select the true model, and vice versa. A good estimator is consistent in both modes. However, these two modes of consistency serve different purposes, and they are often discussed separately.

Penalized least squares and the LASSO

In the literature, the L_1 regularization (e.g., LASSO) has received a great deal of attention because of its sparsity and convexity. Several remarkable contributions have been made in the context of the L_1-penalized least squares problem, which is formulated as a specific form of (1.6):

$$\min_{\boldsymbol{\beta}} \left\{ n^{-1}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}, \tag{1.9}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is the $n \times p$ design matrix with column entries corresponding to the p covariates, $\boldsymbol{\beta}$ is the p-dimensional regression coefficient, and $\|.\|_1$ denotes the L_1 norm. The solution to (1.9) leads to the (sparse) LASSO estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}_\lambda$, which can be used for both parameter estimation and variable (i.e., the column entries of $\boldsymbol{X}$) selection.

Under some regularity conditions on the design matrix, Knight and Fu [2000] established the estimation consistency of $\hat{\boldsymbol{\beta}}_\lambda$, provided $\lambda \to 0$ as $n \to \infty$. Moreover, they showed that the limiting distribution of $\hat{\boldsymbol{\beta}}_\lambda$ has positive probability masses at zero for the coefficients of irrelevant covariates, which provides insightful justification for using LASSO for variable selection.

Zhao and Yu [2006] further characterized the selection consistency of LASSO by studying a stronger but technically more convenient property, sign consistency, which is defined as $P(\text{sgn}(\hat{\boldsymbol{\beta}}_\lambda) = \text{sgn}(\boldsymbol{\beta}^*))$ with $\boldsymbol{\beta}^*$ denoting the true value of $\boldsymbol{\beta}$. In particular, they derived an irrepresentable condition required for the sign (selection) consistency of LASSO. Following their notation, let $\boldsymbol{\beta}^*$ be split into two parts, $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$, with $\boldsymbol{\beta}_2^*$ assumed to be zero. Also, let the design matrix $\mathbf{X}$ be written as $\{\mathbf{X}(1), \mathbf{X}(2)\}$ accordingly. The irrepresentable condition requires that, for some constant positive vector $\boldsymbol{\eta}$ (not depending on n),

$$|\{\mathbf{X}(1)^T\mathbf{X}(1)\}^{-1}\mathbf{X}(1)^T\mathbf{X}(2)| < \mathbf{1} - \boldsymbol{\eta}$$

where $\mathbf{1}$ is a $q \times 1$ vector with all elements equal to 1 and the inequality holds elementwise. The irrepresentable condition can be interpreted as a correlation constraint between the irrelevant covariates $\mathbf{X}(2)$ and the relevant covariates $\mathbf{X}(1)$. Under the irrepresentable condition and additional requirements, Zhao and Yu [2006] showed that the LASSO is sign-consistent for an appropriate choice of λ, even when the number of covariates p diverges with the sample size n at rate $\log p = O(n^{\alpha})$ for some $0 < \alpha < 1$. However, the irrepresentable condition is hard to verify in practice and can become restrictive in high-dimensional situations.

Addressing a slightly different but closely related problem, Meindhausen and Buhlmann [2006] established the selection consistency of LASSO in the context of Gaussian graphical models, under conditions on the design matrix similar to those in Zhao and Yu [2006]. Zou [2006] further provided a necessary condition for the selection consistency of LASSO. In particular, he showed that, under some general assumptions on the design matrix, LASSO is not variable-selection consistent. The selection bias of LASSO is mainly due to the intrinsic difficulty in distinguishing highly correlated covariates. To correct the bias, Zou [2006] proposed an adaptively weighted L_1 penalty, which is defined by

$$\lambda \sum_{j=1}^{p} \omega_j |\beta_j|,$$

where $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_p\}$ denotes a prespecified weight vector. He further suggested $\boldsymbol{\omega} = 1/|\hat{\boldsymbol{\beta}}|$, with $\hat{\boldsymbol{\beta}}$ being some root-n consistent estimator of $\boldsymbol{\beta}$, so that the penalty is decided adaptively by the data: plausible covariates receive a lower penalty. This strategy accelerates the shrinkage of the coefficients of irrelevant covariates, provided the weight $\boldsymbol{\omega}$ is appropriately specified. Under conditions similar to those used in Knight and Fu [2000], the adaptive LASSO has been shown to have both estimation and selection consistency. However, finding such a root-n consistent estimator of $\boldsymbol{\beta}$ might not be straightforward in high-dimensional situations with $p \gg n$.

Penalized likelihood and the nonconvex penalties

Nonconvex regularization methods (e.g., bridge regression and SCAD) have also received considerable research attention. In a seminal paper, Fan and Li [2001] built the theoretical foundation of nonconvex PLMs for feature selection.

In the framework of generalized linear models (GLMs; McCullagh and Nelder [1989]), Fan and Li [2001] showed that there exists a local maximizer of (1.6) that converges to the true value of the model coefficients $\boldsymbol{\theta}^*$ at a rate $O_p(n^{-\frac{1}{2}}+a_n)$ with $a_n = \max\{\phi'_\lambda(|\theta^*_j|) : \theta^*_j \neq 0\}$. This result implies that if we choose an appropriate $\phi_\lambda(.)$ such that $a_n = O(n^{-\frac{1}{2}})$, the corresponding MPLE is root-n consistent for the parameter estimation. In particular, this is the case when the SCAD penalty is used in (1.6) with $\lambda = o(1)$. Moreover, with additional requirements on $\phi_\lambda(.)$ such as concavity in $|\theta|$, Fan and Li [2001] demonstrated that such a root-n estimation-consistent MPLE is also selection consistent. They referred to this feature of some nonconvex PLMs (e.g., SCAD and MCP) as the oracle property; this means that the MPLE consistently identifies influential variables in a model and estimates their coefficients as efficiently as does the MLE based on the true model.

For potential applications with high dimensionality, Fan and Peng [2004] extended the results of Fan and Li [2001] to diverging p cases with $p = o(n^{1/3})$. Recently, Fan and Lv [2011] illustrated the oracle property of nonconvex PLMs even when $\log p = O(n^\xi)$ for some $\xi \in (0, 1)$. In addition, they derived sufficient conditions that guarantee asymptotic equivalence between the global maximizer of the SCAD-penalized likelihood and the oracle estimator. Their results also suggest that L_1-based PLMs generally can not achieve selection and root-n estimation consistency simultaneously, and thus in general they do not have the oracle property.

1.4.3 Tuning strategies

As discussed in Sections 1.4.1 and 1.4.2, the attractive features of PLM depend on an appropriate choice of $\phi_\lambda(.)$. Given a specific form of the penalty function $\phi_\lambda(.)$, we must select a proper tuning parameter λ that controls the amount of regularization. It is well known that excessive regularization may lead to the elimination of important variables, while insufficient shrinkage retains too many irrelevant variables in the model. Hence, the selection of λ is critical for achieving the advantages of PLM in practice.

To address this issue, Tibshirani [1996] used the m-fold cross-validation (CV; Stone [1974]) to select the λ for LASSO. Given the model settings and the notation in Section 1.3.1, this procedure works as follows: First, we divide the full dataset $\boldsymbol{d}$ into T separate sets, say $\{\boldsymbol{d}_t\}$ for $t = 1, \ldots, T$, and then find the MPLE $\hat{\boldsymbol{\theta}}_\lambda(-t)$ of $\boldsymbol{\theta}$ based on $\boldsymbol{d} - \boldsymbol{d}_t$. We choose an optimal λ by minimizing

$$\text{CV}(\lambda) = -\sum_{t=1}^{T} l(\boldsymbol{d}_t, \hat{\boldsymbol{\theta}}_\lambda(-t)) = -\sum_{t=1}^{T} \sum_{i \in \boldsymbol{d}_t} \log f(d_i, \hat{\boldsymbol{\theta}}_\lambda(-t)).$$

The CV-based tuning method provides a λ that yields a model with the "optimal" prediction accuracy.

In the same spirit, Fan and Li [2001] suggested a tuning method based on generalized cross-validation (GCV; Craven and Wahba [1979]), which selects the λ that minimizes

$$\text{GCV}(\lambda) = \frac{-l(\boldsymbol{d}, \hat{\boldsymbol{\theta}}_\lambda)}{n(1 - \tau(s_\lambda)/n)^2},$$

where s_λ denotes the model corresponding to $\hat{\boldsymbol{\theta}}_\lambda$. Compared with the CV method, the GCV method is computationally more convenient, and it has been widely used.

However, from a variable-selection point of view, the GCV-based method has been shown to be inconsistent (Wang et al. [2007]); it tends to select many irrelevant variables. Wang et al. [2007] advocated the use of BIC for tuning a PLM. This approach finds λ by minimizing

$$\text{BIC}(\lambda) = -2l(\boldsymbol{d}, \hat{\theta}_\lambda) + \tau(s_\lambda) \log n.$$

In the context of penalized least squares with finite-parameter settings, Wang et al. [2007] showed that the SCAD estimator with λ chosen by $\hat{\lambda}_{BIC}$ is consistent in variable selection. This result has been extended to the diverging p cases with

$p = o(n)$ in Wang et al. [2009].

In recent work, Chen and Chen [2012] proposed a family of extended Bayesian information criteria (EBIC) under the GLM setup, which is designed for high-dimensional model selection with a sound Bayesian motivation (see Section 2.4.2). EBIC has been found to be an effective tool for choosing the tuning parameter for PLM in situations where $p \gg n$ (Wang [2009], She [2011]). Specifically, the EBIC-based tuning strategy finds the λ that minimizes

$$\text{EBIC}(\lambda) = -2l(\boldsymbol{d}, \hat{\boldsymbol{\theta}}_\lambda)) + \tau(s_\lambda)(\log n + \gamma \log p)$$

for some $0 \leq \gamma \leq 1$. Clearly, for $\gamma = 0$ EBIC reduces to the BIC-based tuning strategy. By choosing $\gamma > 0$, EBIC places more penalties on the model complexity by linking the regularization to the number of candidate variables p. This modification is particularly helpful for selecting an appropriate parsimonious model in ultra-high-dimensional applications where $p \gg n$, such as the genomic example introduced in Section 1.2.1.

1.4.4 Computational strategies

In the last decade, substantial progress has been made in solving optimization problems related to PLMs. Various numerical methods have been proposed for finding the maximizer (i.e., the MPLE) from the penalized likelihood (1.6). In principle, when convex penalties (e.g., L_1 and L_2) are used, the objective function (1.6) is concave and convex optimization algorithms can be conveniently applied. For nonconvex penalties (e.g., SCAD and MCP), however, the task becomes more challenging since (1.6) is no longer concave in general, and the numerical procedure may lead to local maxima. In this subsection, we briefly review several algorithmic contributions for convex and nonconvex PLMs respectively.

LARS and CD

As a fundamental regularization problem, the L_1-penalized least squares (1.9) has been studied extensively. Efficient algorithms have been developed to numerically find the LASSO estimate. In a seminal paper, Tibshirani [1996] treated the numerical problem related to LASSO as a linearly constrained least squares problem, which can be solved by standard quadratic programming techniques. Efron et al. [2004] developed a fast and efficient least angle regression (LARS) procedure for the variable selection, which can be modified to provide the entire solution path for LASSO ($\{\hat{\boldsymbol{\beta}}_\lambda : \lambda > 0\}$).

Following (1.9), the LARS algorithm works roughly as follows. Starting with all the coefficients $\beta_1, \ldots, \beta_p$ equal to 0, LARS finds the covariate $\boldsymbol{x}_l$ that is most correlated with the response $\boldsymbol{y}$. Then, the LARS moves β_l from 0 toward its marginal least squares estimate of $\boldsymbol{y}$ on $\boldsymbol{x}_l$, until some other covariate x_h has the same correlation with the current residual as that of $\boldsymbol{x}_l$. Next, the LARS algorithm moves (β_l, β_h) in the direction defined by their joint least squares estimate of the current residual on $(\boldsymbol{x}_l, \boldsymbol{x}_h)$, until a third covariate has the same correlation with the current residual. The procedure continues in this way until all the covariates have been added to the model. LARS can be modified to provide the exact solution path of LASSO: if any nonzero coefficient reaches zero in a LARS procedure, we drop the corresponding covariate from the model and recompute the current residual. After p steps ($p < n$), we reach the least squares estimate of $\boldsymbol{\beta}$ based on the full model, which corresponds to the LASSO estimate with $\lambda = 0$. The idea of LARS has been extended to efficiently solve the L_1-penalized likelihood problem in the context of GLMs (McCullagh and Nelder [1989]). Because of the computational advantage of the LARS algorithm, LASSO quickly became a popular technique for dimensionality reduction and feature extraction.

Recently, the coordinate-wise descent algorithm (CD) has received attention for its ability to solve the numerical problem related to convex PLMs (Fu [1998], Friedman et al. [2007]). The idea of CD is quite simple: it iteratively optimizes the objective function (1.6) one variable at a time. To many researchers' surprise, this seemingly naive strategy works amazingly well for a wide range of convex regularization problems, including LASSO, the nonnegative garotte (Breiman [1995]), and the elastic net. Taking LASSO as an example, Friedman et al. [2007] demonstrated that CD is competitive with the well-known LARS, and thus it has great potential for high-dimensional problems.

LQA and LLA

In nonconvex PLMs, the irregular shape of the objective function poses a challenge. To address this issue, Fan and Li [2001] suggested locally approximating a nonconvex penalty by a quadratic function (LQA), i.e.,

$$\phi_\lambda(|\theta|) \approx \phi_\lambda(|\theta_0|) + \frac{1}{2}\frac{p_\lambda^{'}(\theta_0)}{|\theta_0|}(\theta^2 - \theta_0^2)$$

for θ close to some initial value θ_0. With the aid of LQA, Newton-type algorithms can be modified to solve the MPLE from (1.6) with a nonconvex penalty. For example, we can use the MLE as the initial value and obtain the MPLE through an updating procedure by

$$\boldsymbol{\theta}^{(k+1)} = \arg\max\left\{ l(\boldsymbol{\theta}) - n\sum_{j=1}^{p}\frac{\phi_\lambda^{'}(|\theta_j^{(k)}|)}{2|\theta_j^{(k)}|}\theta_j^2 \right\}.$$

However, the sequence $\boldsymbol{\beta}^{(k)}$ obtained from LQA may not be sparse for any fixed k and hence is not directly suitable for feature selection. Fan and Li [2001] further suggested setting $\theta_j^{(k)} = 0$ if $|\theta_j^{(k)}|$ is sufficiently small, say $|\theta_j^{(k)}| < \varepsilon_0$ for some tolerance level ε_0, and removing the corresponding covariate from the model. Although it is useful in practice, such an ad hoc step may bring instability to the procedure: once a variable has been deleted, it can never be considered again. Furthermore, the choice of ε_0 has a direct impact on the final selection result. Chossing an appropriate ε_0 might be problematic in applications.

To avoid this drawback, Zou and Li [2008] proposed a one-step algorithm based on local linear approximation (LLA) to the penalty function:

$$\phi_\lambda(|\theta|) \approx p_\lambda(|\theta_0|) + \frac{1}{2}\phi_\lambda^{'}(|\theta_0|)(|\theta| - |\theta_0|),$$

which leads to a similar iteration procedure:

$$\boldsymbol{\theta}^{(k+1)} = \arg\max\left\{ l(\boldsymbol{\theta}) - n\sum_{j=1}^{p}\phi_\lambda^{'}(|\theta_j^{(k)}|)|\theta_j| \right\}.$$

Because this form of LLA shares the common traits of the L_1 penalty, the final estimate has a sparse structure and hence avoids the ad hoc step. More importantly, efficient algorithms developed for convex PLMs (e.g., LARS) can be directly adopted in the updating procedure. Zou and Li [2008] established the convergence of the LLA procedure and further showed that, starting with a root-n consistent estimator, the nonconvex MPLE obtained by LLA has the oracle property after a single iteration. However, obtaining such a "good" initial value may be an issue, especially for the large-p-small-n cases frequently encountered in genetic applications.

Recently, She [2009] proposed an iterative thresholding-based algorithm, which provides a novel approach for nonconcave PLMs. In this strategy, the multivariate nonconcave objective function is transformed into equivalent univariate functions for which simple optimization can be conveniently handled. For the L_1 penalized least squares problem (1.9), the iterative thresholding-based algorithm is identical to the CD procedure. Thus, it shares advantages with the efficient algorithms developed for convex PLMs. We provide a more detailed discussion of the thresholding-based algorithm in Section 4.3.

1.5 Contributions of the dissertation

The PLM has received much attention, and it is applicable to a wide range of statistical models in a variety of scientific areas. In this dissertation, we address new research problems in feature selection arising from several applications of the PLM.

As mentioned in Section 1.2, in computational genomics, the number of candidate features can be much larger than the sample size (ultra-high dimensionality). In such applications, the number of features can be so large that even the computationally attractive PLM is not adequate. A super-efficient procedure to quickly screen a large number of candidate features is therefore essential. Fan and Lv [2008] proposed sure independence screening (SIS), which screens features based on their marginal correlations with the response. It is natural to conjecture that accounting for joint effects among the candidate features will be beneficial for the screening. In this spirit, in Chapter 2 we propose a novel screening approach via the sparsity-restricted maximum likelihood estimator (SMLE) and investigate its performance. The SMLE estimates the high-dimensional model coefficients in a designated low-dimensional subspace and screens features by setting their coefficients to zero. The features passed by SMLE are then subject to a more elaborate selection via the PLM. SMLE incorporates joint effects among features by jointly estimating their model coefficients, and thus it has the potential to provide more reliable screening results than SIS provides. We show that SMLE enjoys the sure screening property of Fan and Lv [2008] in the ultra-high-dimensional GLM setup, and we develop an efficient algorithm for its implementation. In addition, we establish estimation and selection consistency for the SMLE-based PLM and propose the use of EBIC for the tuning parameter selection. The effectiveness of the new methods has been demonstrated in simulation studies.

Regression models are routinely used to analyze survey data; they identify the influential factors for certain social or behavioral indices in a target population. As discussed in the LSCDC example of Section 1.2, when data are collected through survey sampling from a finite population without replacement, they have an intrinsic dependence structure and may not represent the target population. To avoid distorted conclusions, survey weights are usually adopted in the estimation of parameters in regression models based on survey data. Incorporating the survey weights may also be beneficial for variable selection. To investigate this, in Chapter 3 we explore the use of pseudo-likelihood to take account of the survey weights and study a penalized pseudo-likelihood method (PPLM) for the variable selection of survey data. In a joint randomization framework, we prove that the PPLM consistently identifies the influential variables through BIC-based tuning. The finite-sample performance of the approach is assessed via analysis and computer simulations based on data from the hypertension component of the 2009 Survey on Living with Chronic Diseases in Canada. The results show that, compared with the standard PLM, the PPLM helps to avoid biased selection results due to informative sampling and provides protection against model mis-specification.

In market research, finite mixture models are frequently used to depict the heterogeneity of an overall data structure. Selecting the most suitable number of mixture components (the order) is fundamental to these applications. In Chapter 4, we address order selection for finite mixture models, which provides a flexible tool for modeling data from a heterogeneous population. The PLM procedure proposed by

Chen and Khalili [2008] is attractive for order selection. The method fits a high-order mixture model to the data via the penalized likelihood, and the nonsmooth penalty helps to merge close components to achieve a lower order. However, this method requires maximizations over nonsmooth and nonconcave objective functions, which are computationally challenging. The commonly used LQA approach fails to provide a sparse solution, which might lead to unstable selection results.

To tackle this problem, we transform the original multivariate objective functions into a sum of univariate functions and design an iterative thresholding-based algorithm to efficiently solve the sparse maximization without ad hoc steps. We further show the ascent property of the proposed algorithm, i.e., each update to the parameter estimates increases the value of the objective function. This desirable property not only helps to design an appropriate stopping rule in practice but also leads to the convergence of the algorithm. Our simulation studies show that the new algorithm reduces the computational time by approximately 40% in comparison with LQA. To further reduce the computational burden, we propose a revised BIC criterion to select the tuning parameter of Chen and Khalili's method, where the regular BIC penalty on the model complexity is reduced by half. We do this because an extra component in a mixture model usually leads to a single extra degree of freedom in the limiting distribution of the likelihood-based test statistic. We demonstrate the efficiency of the proposed method via numerical studies.

Lastly, in Chapter 5, we summarize the main findings of this dissertation and present several directions for future research. All the technical derivations and proofs are provided in Appendices A–C.

Chapter 2

The Screening-Based PLM in Ultra-High-Dimensional Feature Spaces

2.1 Introduction

High-dimensional datasets with many variables are frequently encountered in modern scientific research (Hastie et al. [2009], Donoho [2000], Fan and Lv [2010]). It is often important to identify the features that influence the response (feature extraction). Examples include detecting the biomarkers responsible for rare diseases and finding the stocks that generate profits in investment portfolios. Regression models are frequently used, and the feature extraction is typically performed by a variable selection procedure.

Traditional selection approaches, such as best subset selection and stepwise regression, can be computationally expensive and instable in the selection process. The PLMs, including LASSO (Tibshirani [1996]), SCAD (Fan and Li [2001]), the elastic net (Zou and Hastie [2005]), and MCP (Zhang [2010]), are now being used as computationally feasible alternatives for variable selection. These approaches are practical and can be selection consistent. However, when the number of features p is much larger than the sample size n (ultra-high-dimensional cases), as is common in genetic studies, the direct use of PLMs is difficult. For PLMs with non-convex penalties (e.g., MCP and SCAD), finding the global maximizer of the corresponding penalized likelihood is computationally challenging. For PLMs with convex penalties (e.g., LASSO and the elastic net), selection consistency may not hold in general. More importantly, the choice of an appropriate tuning parameter for the PLM in ultra-high-dimensional applications has not yet been determined.

To overcome the difficulties of the large-p-small-n situation, Fan and Lv [2008] proposed reducing the number of candidate features to a manageable level before applying a more elaborate selection method. They suggested SIS, which screens the original p features according to their marginal correlations with the response variable. This approach efficiently removes most of the irrelevant features while retaining important features with a high probability (sure screening). Wang [2009] proposed the use of classical stepwise forward regression (FR) for feature screening and established its consistency in the sense of Fan and Lv [2008]. Feature screening techniques simplify the high-dimensional variable selection to a lower-dimensional problem, where the PLM can be conveniently applied.

Motivated by the insights of these methods, in this chapter we propose a novel approach for feature screening via the sparsity restricted maximum likelihood estimator (SMLE). SMLE estimates the high-dimensional model coefficients in a designated low-dimensional subspace and naturally screens features by setting their coefficients to zero. Unlike SIS, SMLE accounts for joint effects between features by jointly estimating their model coefficients. Thus, it has the potential to provide more reliable screening results in practice.

The SMLE approach belongs to a more general class of sparsity-constrained approximation methods that have been widely adopted in wavelet analysis, signal processing, and compressed censoring. Sparsity-constrained methods are frequently used to construct a parsimonious representation (approximation) of high-resolution images/signals for fast transmission and recovery (Donoho [2006], Candès et al. [2006], Blumensath and Davies [2009]). We attempt to investigate the potential of sparsity-constrained methods (i.e., SMLE) for feature screening. In particular, we show that the SMLE enjoys the sure screening property in the sense of Fan and Lv [2008] in the context of high-dimensional GLMs (McCullagh and Nelder [1989]), where the number of covariates p can be considerably larger than the sample size

n. An iterative hard thresholding-based algorithm (IHT; Blumensath and Davies [2008]) can be used to compute SMLE. We establish the consistency of SMLE-based PLM and demonstrate the promising performance of the proposed procedure via numerical studies.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the ultra-high-dimensional GLM and review the SIS-based screening method. In Section 2.3, we investigate the use of SMLE for feature screening and discuss its asymptotic properties; an IHT algorithm is proposed for the implementation. We discuss the SMLE-based PLM in Section 2.4 and assess the finite-sample performance of the method in Section 2.5. Finally, Section 2.6 presents several remarks. The proofs of the theorems are presented in Appendix A.

2.2 Sure screening techniques with ultra-high dimensionality

2.2.1 Model settings and notation

Suppose the data $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$ are collected independently from $(Y, \boldsymbol{x})$, where Y is a response variable and $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is a p-dimensional covariate vector. We postulate a GLM between Y and $\boldsymbol{x}$ as follows. Conditioning on $\boldsymbol{x}$, the distribution of Y is assumed to belong to an exponential family taking the form

$$f(y; \theta) = \exp(\theta y - b(\theta) + c(y)) \tag{2.1}$$

with respect to a σ-finite measure ν and known functions $b(.)$ and $c(.)$. θ is usually called the natural parameter, and it is assumed to take values in a compact space $\Theta \subset R$ such that

$$\int f(y; \theta) d\nu = 1, \quad \text{for any } \theta \in \Theta. \tag{2.2}$$

Given (2.1), it is well known that $\mathrm{E}(Y|\boldsymbol{x}) = \mu = b'(\theta)$ and $\mathrm{Var}(Y|\boldsymbol{x}) = \sigma^2 = b''(\theta)$, where the primes denote derivatives with respect to θ. The covariate $\boldsymbol{x}$ influences the response Y in the form of a linear combination

$$g(\mu) = \boldsymbol{x}^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are the p-dimensional model coefficients and $g(.)$ is a specified link function. For theoretical purposes, it is often convenient to relate the natural parameter θ directly to the covariate $\boldsymbol{x}$, i.e., $\theta = u(\boldsymbol{x}^T\boldsymbol{\beta})$ for $u(.) = (g \cdot \mu)^{-1}$. Of special importance is the canonical link $g = \mu^{-1}$, which leads to a linear expression of θ, i.e., $\theta = \boldsymbol{x}^T\boldsymbol{\beta}$. Several classical GLMs with the canonical link are as follows:

- **Linear regression**: *Suppose $\boldsymbol{x}_i$, y_i has the distribution $N(\mu_i, \sigma^2)$. In this case, we have $\mu_i = \theta_i$ and the natural link $g(\mu_i) = \mu_i$, which implies $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Model (2.1) leads to the classical normal linear regression*

  $$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i \tag{2.3}$$

  where ϵ_i is the normal random error with mean zero and variance σ^2.

- **Logistic regression**: *Suppose $\boldsymbol{x}_i$, y_i has the distribution Bernoulli(p_i). In this case, we have $p_i = \mu_i = \exp(\theta_i)/(1 + \exp(\theta_i))$ and the natural link $g(\mu_i) = \log \mu_i/(1 - \mu_i)$, which implies $p_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))$. Model (2.1) leads to the logistic regression*

  $$\text{logit}\{P(y_i = 1|\boldsymbol{x}_i)\} = \log(\frac{p_i}{1 - p_i}) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{2.4}$$

  which is commonly used for regression analysis with binary dependent variables.

- **Poisson regression**: *Suppose $\boldsymbol{x}_i$, y_i has the distribution Poisson(μ_i). We have $\mu_i = \exp(\theta_i)$ and the natural link $g(\mu_i) = \log(\mu_i)$, which implies $\mu_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$. Model (2.1) leads to the Poisson regression*

  $$\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{2.5}$$

  which is often used for the analysis of count data and multidimensional contingency tables.

With the GLM settings, the effect of each covariate on the response Y is characterized through the size of the corresponding regression coefficient. In applications, when the dimension p is high, it is often believed that only a small number of the covariates in $\boldsymbol{x}$ contribute to the variations in Y, which leads to an idealistic assumption that $\boldsymbol{\beta}$ is sparse. With this sparsity, we can identify influential features by finding the covariates associated with nonzero coefficients. Specifically, let $\boldsymbol{\beta}^*$ be the true sparse coefficients with q nonzero elements, and let s be an arbitrary subset of $\{1, \ldots, p\}$ defining a submodel with covariates $\boldsymbol{x}_s = \{x_j, j \in s\}$ and associated coefficients $\boldsymbol{\beta}_s = \{\beta_j, j \in s\}$. For convenience, we use $\|.\|_0$ to denote the number of nonzero components of an arbitrary vector (i.e., the l_0-norm) and $\tau(s)$ to indicate the size of model s. In particular, we denote the true model by $s^* = \{j : \beta_j \neq 0\}$ with $\tau(s^*) = \|\boldsymbol{\beta}^*\|_0 = q$. We must estimate s^* from $\{1, \ldots, p\}$ by analyzing the data $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. We are interested in solving this problem in ultra-high-dimensional situations where $p \gg n > q$.

2.2.2 Sure independence screening

As mentioned, the high dimensionality of p poses great challenges in searching for s^* using most existing selection methods. A natural idea is to reduce the dimensionality of the feature space p to a manageable level k (say, $k < n$) by a fast and reliable method, such that existing selection methods (e.g., PLM) can be applied to the reduced feature space. This suggests a two-step procedure for feature selection in ultra-high-dimensional situations: a crude feature screening followed by a more careful selection.

To this end, Fan and Lv [2008] developed the SIS framework by ranking the marginal correlations between the covariates and the response. Specifically, let $\boldsymbol{w} = (w_1, \ldots, w_p)^T$ be a p-dimensional vector with component w_j denoting the fitted coefficient obtained from the marginal regressions of the response variable on the jth covariate. For a given screening bound $k < n$, the SIS screens features by selecting

$$\check{s} = \{1 \leq j \leq p : \text{the value of } |w_j| \text{ is among the } k \text{ largest values}\}.$$

Clearly, in this strategy, each feature is used independently as a single covariate to deternine its strength of association with the response. SIS has been widely adopted in genetic studies, where two-sample testing methods are frequently used to detect the genes that differ between case and control groups (Efron [2007], Storey and Tibshirani [2003]).

In the context of a linear model with Gaussian covariates and response, Fan and Lv [2008] showed that

$P(s^* \subset \check{s}) \to 1$ as $n \to \infty$ even when p increases exponentially with n. They refer to this property of SIS as the sure screening property, meaning that a screening method has the ability to retain all the important features with a high probability. This result has been further extended by Fan and Song [2009] to the high-dimensional GLM case where both the response and the covariates can be discrete.

SIS uses only the marginal information of the covariates, and its performance can be unstable in applications. To address this issue, Fan et al. [2009] proposed an iterative SIS (ISIS) procedure, which works as follows. First, ISIS applies SIS to select k_1 features, denoted $\check{s}_1$. Then, it applies the PLM (e.g., SCAD) on $\check{s}_1$ to select a subset $\check{\mathcal{M}}_1 \subset \check{s}_1$. Next, it fits a regression of Y on $\check{\mathcal{M}}_1$ and obtains the corresponding residuals. These residuals are then treated as the new response variable, and SIS is applied to the remaining candidate features $\{x_1, \ldots, x_p\}/\check{\mathcal{M}}_1$ to select k_2 features, say $\check{s}_2$. Then, the PLM is applied to $\check{\mathcal{M}}_1 \cup \check{s}_2$, which gives a new subset $\check{\mathcal{M}}_2$. The procedure continues until there are k features in the current subset.

At each step of ISIS, since the residuals based on $\check{\mathcal{M}}_t$ are uncorrelated with the variables in $\check{\mathcal{M}}_t$, the unimportant variables in $\{x_1, \ldots, x_p\}/\check{\mathcal{M}}_t$ that are highly correlated with the response Y through strong associations with the features in $\check{\mathcal{M}}_t$ are not likely to be selected at the next step. Also, the important features that are marginally weakly correlated with Y because of the presence of the variables in $\check{\mathcal{M}}_t$ should have a chance to be selected. ISIS has great potential to improve SIS, but it has a higher computational cost.

2.3 Variable screening via the sparse-MLE

2.3.1 The sparsity-restricted maximum likelihood estimator

The seminal theory of SIS stimulates us to seek a more effective and efficient method for feature screening in ultra-high-dimensional applications. In particular, we conjecture that incorporating joint information between candidate features will be beneficial. Our investigation starts from the classical likelihood-based inference. Specifically, with the canonical link, the log-likelihood function of $\boldsymbol{\beta}$ is given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \cdot \boldsymbol{x}_i^T \boldsymbol{\beta} - b(\boldsymbol{x}_i^T \boldsymbol{\beta})]. \tag{2.6}$$

Maximizing (2.6) leads to the MLE of $\boldsymbol{\beta}$. Under some regularity conditions, the MLE is a consistent estimate of $\boldsymbol{\beta}^*$ and thus provides information useful for detecting s^*. However, in ultra-high-dimensional settings with $p \gg n$, the classical MLE is not uniquely defined and therefore loses its interpretive ability. In the spirit of regularization, we consider estimating $\boldsymbol{\beta}^*$ in a subspace of R^p with the number of nonzero entries constrained to be less than a given screening bound k. Under the assumption that $\boldsymbol{\beta}^*$ is sparse (i.e., $q < k$), this constrained estimation is expected to retain all the important information carried by $\boldsymbol{\beta}^*$ while setting most of the zero entries of $\boldsymbol{\beta}^*$ exactly to zero, and it therefore provides feature screening.

Motivated by this argument, we now present a new screening method. We propose carrying out the aforementioned estimation via the sparsity-restricted MLE of $\boldsymbol{\beta}$ (SMLE), which is defined by

$$\hat{\boldsymbol{\beta}}_{[k]} = \arg\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) \quad \text{subject to } \|\boldsymbol{\beta}\|_0 \leq k \tag{2.7}$$

for some specified k smaller than n. Clearly, the SMLE $\hat{\boldsymbol{\beta}}_{[k]}$ is constrained to be sparse in its composition, and its nonzero entries correspond to a submodel

$$\hat{s} = \{1 \leq j \leq p : \text{the } j\text{th entry of } \hat{\boldsymbol{\beta}}_{[k]} \text{ is nonzero}\}$$

that yields the highest possible likelihood score within the restricted model sparsity

k. Functionally, the $\hat{\boldsymbol{\beta}}_{[k]}$ screens irrelevant features by setting their coefficients to zero, while retaining the important features in $\hat{s}$ for further selection. Compared with (I)SIS, SMLE naturally accounts for the joint effects between candidate features by jointly estimating their coefficients. Thus, it has the potential to provide more reliable screening results.

The idea of SMLE has similarities with the use of l_0-regularized techniques in image processing, where sparsity-constrained least-squares methods are frequently used to construct parsimonious representations for high-resolution images (Donoho [2006], Blumensath and Davies [2009]). To provide insights into the use of SMLE for feature screening, we first focus on the theoretical aspects and ignore the computational issues at this stage.

As argued in Fan and Lv [2008], a good screening approach should have the ability to remove most of the irrelevant variables while retaining all the relevant ones with a high probability. They refer to this as the sure screening property of a variable screening method. Accordingly, we define SMLE to have screening consistency if

$$P(s^* \subset \hat{s}) \rightarrow 1, \text{ as } n \rightarrow \infty. \tag{2.8}$$

To investigate whether SMLE has screening consistency, we introduce the following notation. For any model s, let

$$\begin{aligned} S(\boldsymbol{\beta}_s) &= \frac{\partial l(\boldsymbol{\beta}_s)}{\partial \boldsymbol{\beta}_s} = \sum_{i=1}^{n} [y_i - b'(\boldsymbol{x}_{is}^T \boldsymbol{\beta}_s)] \boldsymbol{x}_{is}, \\ H(\boldsymbol{\beta}_s) &= -\frac{\partial^2 l(\boldsymbol{\beta}_s)}{\partial \boldsymbol{\beta}_s \partial \boldsymbol{\beta}_s^T} = \sum_{i=1}^{n} b''(\boldsymbol{x}_{is}^T \boldsymbol{\beta}_s) \boldsymbol{x}_{is} \boldsymbol{x}_{is}^T \end{aligned}$$

be the score function and the Hessian matrix of $l(.)$ corresponding to $\boldsymbol{\beta}_s$. For k such that $q < k$, we define

$$\boldsymbol{S}_+^k = \{s : s^* \subset s; \|s\|_0 \leq k\}$$

and

$$\boldsymbol{S}_-^k = \{s : s^* \not\subset s; \|s\|_0 \leq k\}$$

for collections of overfitted models and underfitted models. We investigate the asymptotic properties of $\hat{\boldsymbol{\beta}}_{[k]}$ in the scenario where p, q, k, and $\boldsymbol{\beta}^*$ vary with the sample size n. Also, we assume the following conditions, some of which are purely technical and serve only to provide a theoretical understanding of the new screening method. We do not intend these assumptions to be the weakest possible.

T1 $\log p = O(n^m)$ for some $m > 0$.

T2 There exist nonnegative constants τ_1, τ_2 such that

$$\min_{j \in s^*} |\beta_j^*| \geq w_1 n^{-\tau_1} \quad \text{and} \quad q < k \leq w_2 n^{\tau_2}$$

for some $w_1, w_2 > 0$.

T3 There exist a positive constant c_1 and a corresponding $\delta_1 > 0$ such that for sufficiently large n,

$$\lambda_{min}[n^{-1} H(\boldsymbol{\beta}_s)] \geq c_1$$

for $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta}_s : ||\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*||_2 \leq \delta_1\}$ and $s \in \boldsymbol{S}_+^{2k}$, where $\|.\|_2$ is the Euclidean norm and $\lambda_{min}[.]$ denotes the smallest eigenvalue of a matrix.

T4 There exist positive constants c_2, c_3, c_4, such that

$$\max_{1 \leqslant j \leqslant p} \max_{1 \leqslant i \leqslant n} \left\{ \frac{x_{ij}^2}{\sum_{i=1}^n x_{ij}^2 \sigma_i^2} \right\} \leq c_2 \cdot n^{-1}, \text{ and}$$

$$c_3 \leq n^{-1} \sum_{i=1}^n x_{ij}^2 \sigma_i^2 \leq c_4$$

for any $j \in \{1, \ldots, p\}$.

By condition T1, we assume that p diverges with n up to an exponential rate, which implies that the number of covariates can be substantially larger than the sample size. Condition T2 assumes that $\boldsymbol{\beta}^*$ is sparse and its minimal component does not degenerate too quickly. Condition T3 corresponds to assumptions A4–A5 of Chen and Chen [2012]. It basically requires s^* to stay some distance from incorrect models as n increases. Condition T4 places restrictions on the observed values of

$\mathbf{x}_j$. For a wide range of models, condition T4 holds naturally for random designs on $\boldsymbol{x}$ or for fixed designs with an appropriate rescaling operation.

We now establish the screening consistency (sure screening property) of the SMLE $\hat{\boldsymbol{\beta}}_{[k]}$ via the following theorem.

Theorem 2.1 *Under model (2.1) and conditions T1–T4, if $\tau_1 + \tau_2 < \frac{1}{2}$, we have*

$$P(s^* \subset \hat{s}) \to 1, \text{ as } n \to \infty.$$

See Appendix A for the proof. By Theorem 2.1, we show that, with probability tending to one, SMLE retains all important features in s^* by estimating their model coefficients away from zero. This desirable property of SMLE provides a necessary condition for correctly identifying s^* in the more elaborate selection based on $\hat{s}$, as will be illustrated in Section 2.4.

2.3.2 Implementation

As discussed in the previous subsection, SMLE has the potential to address ultra-high-dimensional feature screening, but this needs to be demonstrated in applications. In principle, numerically finding $\hat{\boldsymbol{\beta}}_{[k]}$ from (2.7) corresponds to a l_0-regularized problem. Such problems have been extensively studied in the area of signal processing, and a number of computational strategies have been developed. Examples include the matching pursuit algorithms (Mallat and Zhang [1993]) and the FOCUSS-based methods (Murray and Kreutz-Delgado [2001]). We find that the hard-thresholding-based algorithms developed under linear models (Blumensath and Davies [2009] are suitable for our needs. In this subsection, we design a modified hard-thresholding procedure in the context of GLM for computing $\hat{\boldsymbol{\beta}}_{[k]}$ in applications.

Specifically, to tackle the problem in (2.7), we first approximate the $l_n(.)$ at a generic point $\boldsymbol{\beta}$ by

$$h_n(\boldsymbol{\gamma}; \boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) + (\boldsymbol{\gamma} - \boldsymbol{\beta})^T s_n(\boldsymbol{\beta}) - \frac{u}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2, \tag{2.9}$$

where $u > 0$ is a scale parameter. The first two terms in (2.9) match the Taylor's expansion of $l_n(\boldsymbol{\gamma})$ at $\boldsymbol{\gamma} = \boldsymbol{\beta}$, and $\frac{u}{2}\|\boldsymbol{\gamma} - \boldsymbol{\beta}\|_2^2$ is introduced as a regularization term. Clearly, we have $l_n(\boldsymbol{\beta}) = h_n(\boldsymbol{\beta}; \boldsymbol{\beta})$, and $h_n(\boldsymbol{\gamma}; \boldsymbol{\beta})$ well approximates $l_n(\boldsymbol{\beta})$ for $\boldsymbol{\gamma}$ close to $\boldsymbol{\beta}$. A key property of $h_n(\boldsymbol{\gamma}; \boldsymbol{\beta})$ is that it is additive in the components of $\boldsymbol{\gamma}$, so that the maximization of $h_n(\boldsymbol{\gamma}; \boldsymbol{\beta})$ over $\boldsymbol{\gamma}$ can be conveniently carried out.

With the aid of (2.9), we then propose the following iterative procedure to solve (2.7)

$$\boldsymbol{\beta}^{(t+1)} = \arg\max_{\boldsymbol{\gamma}} h_n(\boldsymbol{\gamma}; \boldsymbol{\beta}^{(t)}) \quad \text{subject to } \|\boldsymbol{\gamma}\|_0 \leq k. \tag{2.10}$$

For each iteration step, the regularization term in $h_n(.)$ prevents the maximizer from being far from the current estimate of $\boldsymbol{\beta}$. This feature further guarantees the convergence of (2.10) to a local maximum of $l_n(.)$ (subject to the sparsity constraint).

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{X} = (\boldsymbol{x}_i, \ldots, \boldsymbol{x}_n)^T$. Because of the additivity of $h_n(\boldsymbol{\gamma}; \boldsymbol{\beta})$ in $\boldsymbol{\gamma}$, the optimization in (2.10) takes a unified specific form:

$$\min_{\boldsymbol{\gamma}} \frac{1}{2} \left\| \boldsymbol{\gamma} - u^{-1}[u\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T b'(\boldsymbol{X}\boldsymbol{\beta})] \right\|_2^2 \quad \text{subject to } \|\boldsymbol{\gamma}\|_0 \leq k. \tag{2.11}$$

Obviously, if the sparsity constraint is ignored, the solution to (2.11) is the typical least squares estimate

$$\tilde{\boldsymbol{\gamma}} = \boldsymbol{\beta} + u^{-1}\boldsymbol{X}^T[\boldsymbol{y} - b'(\boldsymbol{X}\boldsymbol{\beta})],$$

which corresponds to a zero loss in the objective function. Thus, the constrained minimum of (2.11) is achieved by choosing the k largest (in absolute value) components of $\tilde{\boldsymbol{\gamma}}$. Consequently, the solution to (2.11) is given by

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{H}(\tilde{\boldsymbol{\gamma}}; k) = \left[H(\tilde{\gamma}_1; |\tilde{\boldsymbol{\gamma}}|_{[k]}), \ldots, H(\tilde{\gamma}_p; |\tilde{\boldsymbol{\gamma}}|_{[k]})\right]^T,$$

where $|\tilde{\boldsymbol{\gamma}}|_{[k]}$ is the kth largest component in $|\tilde{\boldsymbol{\gamma}}|$ and

$$H(\gamma; r) = \begin{cases} \gamma, & \text{if } |\gamma| > r \\ 0, & \text{if } |\gamma| \leq r \end{cases}$$

is the hard thresholding function.

Therefore, the iteration (2.10) can be re-written as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{H}(\boldsymbol{\beta}^{(t)} + u^{-1}\boldsymbol{X}^T[\boldsymbol{y} - b'(\boldsymbol{X}\boldsymbol{\beta}^{(t)})]; k). \tag{2.12}$$

It can be seen that (2.12) is a simple thresholding-based iterative procedure which does not involve complicated operations (such as matrix inversion). This advantage makes (2.12) suitable for high-dimensional computing. Unlike the typical thresholding method, (2.12) adaptively performs hard thresholding on the current update such that each $\boldsymbol{\beta}^{(t)}$ satisfies the sparsity constraint i.e., $\|\boldsymbol{\beta}^{(t)}\|_0 \leq k$. Following Blumensath and Davies [2009], we refer to (2.12) as the iterative hard-thresholding (IHT) procedure.

We now show that the sequence $\{\boldsymbol{\beta}^{(t)}\}$ based on IHT has a property analogous to that of typical thresholding-based methods: it stepwise increases the value of $l(.)$. This increment property further ensures the convergence of IHT to a local maximum of $l(.)$ within the feasible region. For convenience of illustration, we introduce the following condition

T3$'$ $\lambda_{min}[n^{-1}H(\boldsymbol{\beta}_s)] > 0$ for any s with $\tau(s) \leq k$.

It can be seen that condition T3$'$ is analogous to T3, which requires the strict concavity of $l_n(\boldsymbol{\beta}_s)$ over models of size no larger than k. This condition is purely technical and might be further weakened. We focus on providing a theoretical understanding of the IHT and leave this issue to future research. We now justify the convergence of IHT by the following theorem.

Theorem 2.2 *Given the settings and notation introduced earlier, assume that $b(.)$ in (2.1) is twice continuously differentiable. Let $\{\boldsymbol{\beta}^{(t)}\}$ be the sequence defined by (2.12). Denote by ρ_1 the maximum eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$ and*

$$\rho^{(t)} = \max_i \sup_{0<\alpha<1} b''(\alpha \boldsymbol{x}_i^T \boldsymbol{\beta}^{(t+1)} + (1-\alpha)\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)}).$$

If $u > \rho_1 \rho^{(t)}$, then

$$l(\boldsymbol{\beta}^{(t+1)}) \geq l(\boldsymbol{\beta}^{(t)}).$$

Moreover, if condition T3$'$ holds, then $\{\boldsymbol{\beta}^{(t)}\}$ converges to a local maximum of $l_n(\boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_0 \leq k$.

See Appendix A for the proof. Theorem 2.2 implies that, for an appropriate scale parameter u, the IHT necessarily converges. In our simulations, we choose u adaptively at each step according to the value of $\boldsymbol{\beta}^{(t)}$ to guarantee the monotonicity of $l(\boldsymbol{\beta}^{(t)})$. Since $\boldsymbol{\beta}^{(t)}$ may lead to a local maximum, multiple initial values are often used in the hope that the global maximum is not missed. For the linear model, when the IHT starts with $\boldsymbol{\beta}^{(0)} = 0$, its first iteration corresponds to the SIS step based on the marginal information of features. This suggests that zero might be a reasonable initial choice for IHT. In practice, the LASSO estimates of $\boldsymbol{\beta}$ are also convenient for setting $\boldsymbol{\beta}^{(0)}$. The efficiency of IHT has been observed in numerical studies.

2.3.3 Numerical assessment

Model Type	Setup	$(p,\ n)$	SIS	ISIS	FR	LASSO	SMLE
	1	(10000, 200)	.22	.94	1.00	.98	.99
Linear	2	(5000, 120)	.58	.63	.34	.89	.78
	3	(1000, 100)	.01	.73	.88	.28	.99
	1	(1000, 400)	.94	1.00	.- -	.99	.99
Logistic	2	(1000, 400)	.11	.89	.- -	.85	.97
	3	(1000, 400)	.02	.61	.- -	.17	.77
	1	(1000, 200)	.07	.94	.- -	.67	.97
Poisson	2	(1000, 200)	.01	.84	.- -	.39	.94
	3	(1000, 200)	.00	.54	.- -	.01	.93

Table 2.1: Frequencies of covering the true model s^* based on different screening methods

To provide a quick assessment of SMLE-based screening (2.7), we briefly summarize some simulation results based on our numerical studies, which will be presented in more detail in Section 2.5. In this subsection, we focus on showing the numerical performance of the screening methods; we give detailed descriptions of the simulation setups in Section 2.5.

We conduct simulation studies in three different modeling contexts: linear regression, logistic regression, and Poisson regression. In each case, we examine SMLE for three different setups with specific correlation structures among the candidate features. We evaluate the screening method by measuring the frequency with which the resulting model includes all the features in the true model, i.e., we measure the ability to correctly screen the irrelevant features. For the linear model, we compare SMLE with four other screening methods: SIS, ISIS, FR, and LASSO. We do not include FR for the logistic and Poisson models because of its high computational expense. For a fair comparison of the methods, the size of the screened model k is the same for each screening procedure.

Table 2.1 summarizes the coverage frequencies of the true model s^* for different screening methods based on 1000 repetitions. The simulation results reveal that SMLE is competitive with other popular screening methods. The advantage of SMLE-based screening is especially clear for the third correlation setup in the logistic and Poisson models, where both SIS and LASSO miss important features. Compared with ISIS, SMLE achieves higher coverage frequencies at a lower computational cost. A more detailed discussion of the simulation results will be given in Section 2.5.

2.4 Screening-based PLM selection procedure

We have proposed SMLE, which efficiently screens the irrelevant features in an ultra-high-dimensional model while retaining the important features with an overwhelming probability. Given such a feature-screening technique, s^* can be conveniently estimated through a two-step procedure: First, we shrink the full model $\{1, \ldots, p\}$ to a refined submodel $\hat{s}$ of size $\tau(\hat{s}) \leq k < n$, and then we apply the PLM to $\hat{s}$ to identify s^*. We refer to SMLE followed by the PLM as SMLE-PLM. Because many irrelevant covariates are removed from the original feature space, the search for s^* in the PLM step is dramatically narrowed. Sure screening makes it feasible to do feature selection for ultra-high-dimensional problems and dramatically speeds up the selection process. In this section, we present further discussion of SMLE-PLM.

2.4.1 Consistency of SMLE-PLM

As discussed in Section 1.4.2, some PLMs have both estimation and selection consistency for a wide range of applications (Fan and Li [2001]). We now explore whether these desirable properties hold for SMLE-PLM. Unlike the direct implementation of PLM, SMLE-PLM is a two-stage procedure. Given the settings and notation in Section 2.2, SMLE-PLM estimates $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$, found by maximizing

$$Q(\boldsymbol{\beta}_{\hat{s}}) = l(\boldsymbol{\beta}_{\hat{s}}) - n \sum_{j \in \hat{s}} \phi_\lambda(|\beta_j|), \tag{2.13}$$

where $\hat{s}$ is the k-dimensional submodel obtained from SMLE and $\phi_\lambda(.)$ is a specified penalty function. As in typical PLMs, an appropriate choice of $\phi_\lambda(.)$ leads to a sparse $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$, which further identifies important covariates based on $\hat{s}$.

We study the large-sample properties of $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$ under conditions T1–T4 given in Section 2.3.1. For the asymptotic analysis, we associate λ with n and consider a penalty function sequence $\phi_\lambda(.)$ that satisfies the following properties:

P1 For any $\lambda > 0$, $\phi_\lambda(|\theta|) \geq 0$ with $\phi_\lambda(0) = 0$.

P2 For any $\lambda > 0$, $\phi_\lambda^{'}(|\theta|) = \partial \phi_\lambda(|\theta|)/\partial|\theta|$ exists and is continuous for $|\theta| \in (0, +\infty)$.

P3 There exist positive constants τ_3 and w_3, such that $\phi_\lambda^{'}(|\theta|) \leq w_3 n^{-\tau_3}$ for $|\theta| \geq 0.5 w_1 n^{-\tau_1}$.

Properties P1–P2 specify the shape and smoothness of the penalty function, and are generally required for PLMs (e.g., Fan and Lv [2011]). Property P3 further restricts the sequence of penalties by setting an upper bound on the derivatives. For a specific choice of $\phi_\lambda(.)$ (e.g., L_1 or SCAD), P3 corresponds to assuming an appropriate asymptotic order for the tuning parameter λ.

In general, it is difficult to study the global maximizer of (2.13) analytically without the concavity of (2.13). As is common in the PLM literature, we study the behavior of local maximizers. We first establish estimation consistency for SMLE-PLM via the following theorem:

Theorem 2.3 *Under conditions T1–T4, if* $\tau_1 + \tau_2 < \frac{1}{2}$ *and* $\tau_3 > \tau_1 + \frac{\tau_2}{2}$, *then there exists a local maximizer* $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$ *of (2.13) with* $\phi_\lambda(.)$ *satisfying P1–P3, such that*

$$\|\hat{\boldsymbol{\beta}}_\lambda(\hat{s}) - \boldsymbol{\beta}^*\|_2 = O_p(n^{-\upsilon})$$

for some $\upsilon \in (\tau_1, \min\{\frac{1}{2} - \tau_2, \tau_3 - \frac{\tau_2}{2}\})$.

See Appendix A for the proof. Theorem 2.3 shows that, for an appropriate choice of penalty $\phi_\lambda(.)$, SMLE-PLM consistently estimates the model parameters in the ultra-high-dimensional GLM setup. For the situation where the number of influential features q does not diverge with n, we have $\tau_2 = 0$ in condition T2. In this case, SMLE-PLM achieves root-n consistency if $\tau_3 > 0.5$ in requirement P3. In particular, when the L_1 penalty is used (i.e., $\phi_\lambda(|\theta|) = \lambda|\theta|$), this result implies that root-n consistency is achieved if we set $\lambda = o(n^{-\frac{1}{2}})$.

Moreover, we establish the selection consistency of SMLE-PLM with the additional condition:

T5 There exist a positive constant c_5 and a corresponding $\delta_2 > 0$ such that for sufficiently large n,

$$\frac{1}{n}|\frac{\partial l_n(\boldsymbol{\beta}_s)}{\partial \beta_j} - \frac{\partial l_n(\boldsymbol{\beta}_s^*)}{\partial \beta_j}| \leq c_5 \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2$$

for any $j \in s$, $s \in \boldsymbol{S}_+^k$ and $\boldsymbol{\beta}_s \in \{\boldsymbol{\beta}_s : \|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \leq \delta_2\}$.

Condition T5 is purely technical and basically requires Lipschitz continuity of the score function in a neighborhood of $\boldsymbol{\beta}^*$. Also, we need an extra requirement on the order of the penalty function sequence as follows:

P4 There exist positive constants τ_4 and w_4 such that $\phi_\lambda^{'}(|\theta|) \geq w_4 n^{-\tau_4}$ for $|\theta| < 0.5 w_1 n^{-\tau_1}$.

Theorem 2.4 *Under conditions T1–T5, if* $\tau_1 + \tau_2 < 0.5$ *and* $\tau_3 - \tau_2 > 0.5 - 1.5\tau_2 > \tau_4$*, then there exists a local maximizer* $\hat{\boldsymbol{\beta}}_\lambda(\hat{s}) = (\hat{\beta}_{1\lambda}(\hat{s}), \ldots, \hat{\beta}_{k\lambda}(\hat{s}))^T$ *of (2.13) with* $\phi_\lambda(.)$ *satisfying P1–P4, such that*

$$P\{\hat{\beta}_{j\lambda}(\hat{s}) = 0, \textit{ for } j \in \hat{s} \setminus s^*\} \to 1.$$

See Appendix A for the proof. Theorem 2.4 implies that, for an appropriate choice of $\phi_\lambda(.)$, SMLE-PLM consistently selects s^* even when the number of features diverges exponentially with the sample size. Appropriate choices of $\phi_\lambda(.)$ include popular nonconvex penalties such as SCAD or MCP as special cases. In particular, for the fixed-q situation, the selection consistency of SMLE-SCAD is implied if $\lambda = o(1)$ and $\lambda n^{\frac{1}{2}} \to \infty$.

The results of theorems 2.3 and 2.4 are not restricted to SMLE-PLM but also apply to a general screening-based PLM procedure with the sure screening property (e.g., SIS or FR).

2.4.2 Tuning with EBIC

As for standard PLMs, given the form of the penalty function $\phi_\lambda(.)$, we must specify the tuning parameter λ for the implementation of SMLE-PLM. The commonly used GCV or BIC tuning methods (Section 1.4.3) are designed for the situation where $p < n$ and may not be suitable for situations where $p \gg n$. To address this issue, we choose the λ that minimizes the following EBIC criterion (Chen and Chen [2008]),

$$\text{EBIC}(s_\lambda) = -2l(\hat{\boldsymbol{\beta}}_{s_\lambda}) + \tau(s_\lambda)(\log n + \gamma \log p), \quad 0 \leq \gamma \leq 1, \tag{2.14}$$

where s_λ denotes the model corresponding to $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$ (2.13) and $\hat{\boldsymbol{\beta}}_s$ is the MLE based on model s. With $\gamma = 0$, EBIC reduces to the standard BIC tuning strategy (Wang et al. [2007]).

The EBIC is designed for high-dimensional model selection with a sound Bayesian motivation. Let the model space be partitioned into subclasses according to the number of covariates that a model contains. Let S_j for $0 \leq j \leq p$ be the subclass of models containing j covariates and $\tau(S_j)$ be the size of S_j. Let us assign each model in subclass S_j an equal probability, i.e., $P(s|S_j) = 1/\tau(S_j)$ for any $s \in S_j$. Then, instead of assigning probabilities $P(S_j)$ proportional to $\tau(S_j)$ as in BIC, EBIC sets $P(S_j)$ to be proportional to $\tau(S_j)^{1-\gamma}$ for some γ between 0 and 1. Thus, in EBIC the prior probability $P(s)$ for $s \in S_j$ is set to be proportional to $\tau(S_j)^{-\gamma}$. Compared with the constant prior used in BIC, this assignment substantially reduces the high prior in models with a large number of covariates, and hence may be suitable for large p. Chen and Chen [2012] have shown that EBIC is selection consistent in the GLM context even when $\log p = O(n^m)$.

Theorem 2.3 shows that there exists a proper sequence of $\phi_\lambda(.)$ such that

$P(s_\lambda = s^*) \to 1$. The EBIC can help to tune $\phi_\lambda(.)$ for the implementation of screening-based PLM. In our simulation studies, we set $\gamma = 0.5$ for EBIC as suggested by Chen and Chen [2012].

2.5 Numerical studies

We assess the finite sample performance of the SMLE-PLM via simulation studies. In particular, we are interested in knowing how SMLE compares with other popular screening methods.

We must consider many factors in order to obtain a relatively complete picture of the new method. Recall that our general goal is to use a GLM to explain the variation in the response variable Y through a number of covariates (features) selected from a large number of candidates. The correlation structure between these covariates can have a strong effect on the performance of screening-based methods. Also, their performance may vary depending on the model to which they are applied. In addition, there are implementation issues to address. Last but not least, we need measures to evaluate the performance of the different methods. In the next subsections we discuss the simulation settings. Some of the simulation results have already been revealed and additional ones will be presented.

2.5.1 General settings

We examine the methods in three different modeling contexts: linear regression, logistic regression, and Poisson regression, as described in Section 2.2.1. For the linear model, we compare screening-based LASSO and SCAD on five screening methods: SIS, ISIS, FR, LASSO, and SMLE. We do not include FR for the logistic and Poisson models because of its computational cost.

Correlation structure

For each model we consider three different correlation structures of the features $x_1, \dots, x_p$, so there are a total of nine combinations. The correlation structures are as follows:

Setup 1: $x_1, \dots, x_p$ are independent and identically distributed $N(0, 1)$ random variables.

Setup 2: $x_1, \ldots, x_p$ are joint Gaussian, marginally $N(0,1)$, with $\text{cov}(x_j, x_{j-1}) = 2/3$, $\text{cov}(x_j, x_{j-2}) = 1/3$ for $j \geq 3$, and $\text{cov}(x_j, x_h) = 0$ if $|j - h| \geq 3$.

Setup 3: $x_1, \ldots, x_p$ are joint Gaussian, marginally $N(0,1)$, with $\text{cov}(x_j, x_h) = 0.15$ for $j, h \in s^*$ and $\text{cov}(x_j, x_h) = 0.3$ for j or $h \in \{1, \ldots, p\} \setminus s^*$.

Case 1 is the ideal independence structure, which is the most straightforward for variable selection. In Case 2, we consider a moving average type correlation structure, where features are strongly correlated for order distances less than three. This type of correlation is commonly used to model features with a natural order. In Case 3, we have a compound correlation structure such that every irrelevant feature has equal correlation with the relevant features. We therefore expect the variable selection in Case 3 to be more challenging.

Implementation issues

In our simulation studies, the screening methods are implemented as follows.

For ISIS, we use the R function **GLMvanISISscad** in the **SIS** package, with a maximum of five ISIS loops. The number of relevant features retained in each loop is decided by SCAD with the AIC-based tuning method; see Fan et al. [2009]. For SMLE, we use the proposed IHT algorithm. In particular, we choose ten LASSO estimates with sparsity varying from $n - 1$ to k as different initial values for IHT. The estimate from IHT that maximizes the likelihood is then treated as the SMLE. For comparison purposes, we also report the performance of LASSO when it is applied directly to the full model. In particular, we apply the R function **glmnet** to identify a sequence of ordered features with a prespecified size k. Variables in the sequence are considered to be important (after screening), and PLMs are used for the further selection.

For each model, we use NONE to represent the model without further variable selection. We use the notation LASSO and SCAD to represent, respectively, the models selected by LASSO and SCAD with the EBIC tuning strategy. To facilitate the computing process, we use the built-in functions in the R software packages **glmnet** and **SIS** to compute the estimators of LASSO and SCAD respectively.

Theorem 2.1 shows that when the model size k has a certain asymptotic order, SMLE consistently includes all the important features. However, in practice, a specific choice of k must be made in the implementation. Intuitively, a larger choice of k may increase the probability that a screening method includes all the relevant features. However, if the screened model has many irrelevant variables the final selection may be more difficult. In our simulations, we set $k = [n/\log n]$ for the linear model, $k = [n/4\log n]$ for the logistic regression model, and $k = [n/2\log n]$ for the Poisson regression model. These model-based choices of k were recommended by Fan et al. [2009] and worked satisfactorily in our simulation examples. In fact, the performance of screening-based PLMs are quite robust to a wide range of sensible values for k. Since our goal is to compare different screening methods, we treat these model-based k values as benchmarks for the comparison.

Assessment of performance

We assess the performance of the screening methods based on $T = 1000$ simulation replications. Specifically, let $\hat{s}_t$ denote the model selected in the jth replication by a particular method (e.g., SMLE-None). The coverage probability of that method is computed by $T^{-1}\sum_{t=1}^{T} I(s^* \subset \hat{s}_t)$, which measures its ability to discover all relevant features. We characterize the model selectivity of each method in terms of the positive selection rate (PSR) and the false discovery rate (FDR), which are defined as follows:

$$\text{PSR} = \frac{\sum_{t=1}^{T} \tau(s^* \cap \hat{s}_t)}{T\tau(s^*)}, \quad \text{FDR} = \frac{\sum_{t=1}^{T} \tau(\hat{s}_t/s^*)}{T\tau(\hat{s}_t)}.$$

The PSR and FDR depict two different aspects of the selection result: a high PSR indicates that most of the important variables have been identified, while a low FDR indicates that only a few irrelevant variables have been selected. We also report the average number of features included in $\hat{s}_t$, as well as the proportion of times that s^* is perfectly identified with no extraneous variables and no missed variables (the correct selection rate, CSR).

In the following subsections, we further specify the model parameter settings and discuss the simulation results.

2.5.2 Linear regression

Parameter settings

In this example, the data $(y_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$ were generated as independent copies from $(Y, \boldsymbol{x})$ according to the linear model (2.3)

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

where ϵ is a normal random error with mean zero and variance σ^2. The model parameters used in each of the three correlation setups are as follows:

Setup 1: $(n, p, \sigma) = (200, 10000, 3)$, and s^* is randomly chosen from $\{1, \ldots, p\}$ with $\|s^*\|_0 = 8$. For $j \in s^*$, the β_j values are generated independently from $U(4 \log n \sqrt{n} + |Z|)$, where U is a binary random variable with $P(U = 1) = 0.6$ and $P(U = -1) = 0.4$ and Z is a standard normal random variable. For $j \notin s^*$, the β_j values are set to zero.

Setup 2: $(n, p, \sigma) = (120, 5000, 5)$, and $s^* = \{1, 3, 5, 7, 9\}$. $(\beta_1, \beta_3, \beta_5, \beta_7, \beta_9) = (5, 3.5, 2.8, 2.5, 2.2)$ and $\beta_j = 0$ for $j \notin s^*$.

Setup 3: $(n, p, \sigma) = (100, 1000, 1)$, and $s^* = \{1, 2, 3, 4\}$. $\beta_j = 2.5$ for $1 \le j \le 4$ and $\beta_j = 0$ for $j > 4$.

The settings in setup 1 are borrowed from example 1 of Wang [2009], where both s^* and $\boldsymbol{\beta}$ are generated randomly. For the settings in setup 2, we choose s^* to ensure non-negligible correlations among all the relevant features. For Case 3, the most challenging situation, we fix s^* to be the set of the first four features and set the coefficients for all the relevant feature to 2.5. The σ values for all three cases were chosen after pilot studies to give an appropriate signal-to-noise ratio.

To evaluate the prediction accuracy of each model, we independently generated testing data $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$ with the same sample size as the training data. For each $\hat{s}_t$ and its associated estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}_t$, we calculated the corresponding relative prediction error by

$$\frac{1}{T} \sum_{t=1}^{T} \left\{ 1 - \frac{\sum_i (\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}_{s^*})^2}{\sum_i (\tilde{y}_i - \tilde{\boldsymbol{x}}_i^T \hat{\boldsymbol{\beta}}_t)^2} \right\},$$

where $\hat{\boldsymbol{\beta}}_{s^*}$ is the least squares estimate based on the true model s^*. For convenience of comparison, we set $\hat{\boldsymbol{\beta}}_t$ to the least squares estimate on $\hat{s}_t$ for a screening method without further selection, and we set $\hat{\boldsymbol{\beta}}_t$ to the corresponding shrinkage estimate for a screening method followed by a PLM (i.e., LASSO or SCAD).

Results

The simulation results are summarized in Tables 2.2–2.4. For setup 1, most screening methods performed very well in terms of including all the relevant features in the model. This is indicated by the high coverage probabilities. However, SIS, which screens features based on marginal correlations, performed poorly because of the high spurious correlations caused by the ultra-high dimensionality.

The drawback of FR is observed in setup 2, where the relevant features are correlated with each other. If one of two highly correlated features is included in an FR procedure, the likelihood of including the other is very small because of its weak correlation with the current residual. Consequently, FR and its associated PLMs performed poorly in this situation. Meanwhile, SIS improved because of the aggregated marginal correlations, but this structure does not encourage further improvements via ISIS. The performance of LASSO and SMLE remained satisfactory.

Lastly, in setup 3, we see that the strong collinearity among the features greatly deteriorates the performance of SIS and LASSO. In terms of the coverage probability, ISIS significantly improved on SIS by almost 72%, while FR improves on LASSO by 88%. In comparison, SMLE did amazingly well in this challenging situation by achieving coverage probabilities as high as 99%.

From Tables 2.2 to 2.4, the performance of LASSO and SCAD after the screening step was similar for setups 1 and 2. SCAN had the better performance for setup 3 in terms of lower FDR and higher CSR. Also, the final model fitted by SCAD consistently yielded a lower prediction error than that fitted by LASSO. This might be due to the excessive shrinkage of the LASSO estimator for large coefficients (see Fan and Li [2001]). In particular, when LASSO and SCAD were used after SMLE in setup 3, the difference in the relative prediction error of their selected (fitted) models was as large as 91%.

Screening		Coverage				Ave.	Relative
Method	PLM	prop.	PSR	FDR	CSR	model size	pred. error
SIS	None	.22	.85	.82	.00	38.0	.52
	LASSO	.21	.84	.11	.15	7.7	.62
	SCAD	.22	.85	.12	.21	7.8	.36
ISIS	None	.94	.99	.79	.00	38.0	.46
	LASSO	.84	.98	.10	.40	8.9	.58
	SCAD	.92	.99	.07	.51	8.6	.25
FR	None	1.00	1.00	.79	.00	38.0	.55
	LASSO	.99	.99	.07	.59	8.7	.44
	SCAD	.99	.99	.09	.37	8.8	.10
LASSO	None	.98	.99	.77	.00	35.6	.38
	LASSO	.85	.98	.11	.39	9.0	.59
	SCAD	.96	.99	.08	.46	8.8	.31
SMLE	None	.99	1.00	.79	.00	38.0	.48
	LASSO	.91	.99	.09	.47	8.8	.56
	SCAD	.98	.99	.07	.49	8.6	.13

Table 2.2: Simulation results for linear regression, setup 1

Screening		Coverage				Ave.	Relative
Method	PLM	prop.	PSR	FDR	CSR	model size	pred. error
SIS	None	.58	.91	.82	.00	25.0	.35
	LASSO	.32	.83	.09	.23	4.6	.42
	SCAD	.26	.78	.15	.17	4.7	.26
ISIS	None	.63	.92	.81	.00	25.0	.42
	LASSO	.38	.85	.10	.24	4.8	.41
	SCAD	.25	.78	.21	.13	5.0	.30
FR	None	.34	.78	.84	.00	25.0	.60
	LASSO	.29	.76	.23	.17	5.1	.33
	SCAD	.19	.73	.28	.08	5.2	.25
LASSO	None	.89	.97	.79	.00	23.4	.35
	LASSO	.44	.86	.09	.26	4.8	.41
	SCAD	.31	.80	.17	.19	4.9	.31
SMLE	None	.78	.95	.81	.00	25.0	.44
	LASSO	.46	.87	.09	.29	4.8	.39
	SCAD	.25	.79	.19	.11	5.0	.28

Table 2.3: Simulation results for linear regression, setup 2

Screening Method	PLM	Coverage prop.	PSR	FDR	CSR	Ave. model size	Relative pred. error
SIS	None	.01	.33	.94	.00	22.0	.89
	LASSO	.01	.23	.91	.00	9.8	.95
	SCAD	.01	.24	.89	.00	9.7	.94
ISIS	None	.73	.85	.84	.00	22.0	.58
	LASSO	.52	.70	.66	.01	8.9	.87
	SCAD	.73	.79	.29	.52	5.8	.25
FR	None	.88	.89	.84	.00	22.0	.59
	LASSO	.78	.86	.53	.02	7.9	.83
	SCAD	.88	.89	.19	.57	5.1	.11
LASSO	None	.28	.64	.87	.00	20.7	.69
	LASSO	.03	.32	.86	.00	9.8	.95
	SCAD	.28	.49	.63	.27	8.2	.68
SMLE	None	.99	1.00	.82	.00	22.0	.52
	LASSO	.52	.82	.61	.01	8.9	.92
	SCAD	.99	.99	.07	.71	4.4	.01

Table 2.4: Simulation results for linear regression, setup 3

2.5.3 Logistic regression

Parameter settings

In our second example, the generic response Y follows a Bernoulli distribution with success probability π satisfying

$$\log(\frac{\pi}{1-\pi}) = \boldsymbol{x}^T\boldsymbol{\beta}.$$

Thus, data generated as independent pairs from $(Y, \boldsymbol{x})$ satisfy a logistic regression model. The parameters used for the three correlation setups are as follows:

Setup 1: s^* is randomly chosen from $\{1, \dots, p\}$ with $\|s^*\|_0 = 8$. For $j \in s^*$, the β_j values are generated independently from $U(4\log n/\sqrt{n} + |Z|/4)$, where U is a binary random variable with $P(U = 1) = 0.5$ and $P(U = -1) = 0.5$ and Z is a standard normal random variable. For $j \notin s^*$, the β_j values are set to zero.

Setup 2: $s^* = \{1, 3, 5, 7, 9\}$. $(\beta_1, \beta_3, \beta_5, \beta_7, \beta_9) = (2, -1.8, 1.6, -1.4, 1.2)$ and

$\beta_j = 0$ for $j \notin s^*$.

Setup 3: $s^* = \{1, 2, 3, 4\}$. $\beta_j = 1.5$ for $1 \leq j \leq 4$ and $\beta_j = 0$ for $j > 4$.

Because of the lack of information in a binary response, we choose $n = 400$ and $p = 1000$ for all three cases. The coefficients are a rescaled version of those in the linear example. The size of the screened model was set to $k = [n/4 \log n]$ for all the screening methods.

Results

The results are shown in Tables 2.2–2.4, where the prediction accuracy of each selected model was evaluated by the proportion of correct predictions based on independent testing data. Again, since the features are independent, all the methods performed well for setup 1. SMLE and its associated PLMs have the best performance in setups 2 and 3, where the correlation structure is more complex. LASSO continues to suffer from the collinearity between the features, particularly in setup 3.

Screening Method	PLM	Coverage prop.	PSR	FDR	CSR	Ave. model size	Prediction accuracy
SIS	None	.94	.99	.53	.00	17.0	.84
	LASSO	.94	.99	.02	.84	8.1	.84
	SCAD	.94	.99	.02	.84	8.1	.86
ISIS	None	1.00	1.00	.53	.00	17.0	.83
	LASSO	.99	1.00	.03	.81	8.2	.84
	SCAD	.99	1.00	.05	.63	8.5	.86
LASSO	None	.99	1.00	.49	.00	15.7	.83
	LASSO	.99	.99	.02	.84	8.2	.84
	SCAD	.99	1.00	.04	.68	8.4	.86
SMLE	None	.99	1.00	.53	.00	17.0	.83
	LASSO	.99	.99	.02	.82	8.2	.84
	SCAD	.99	1.00	.04	.67	8.4	.86

Table 2.5: Simulation results for logistic regression, setup 1

Screening Method	PLM	Coverage prop.	PSR	FDR	CSR	Ave. model size	Prediction accuracy
SIS	None	.11	.74	.78	.00	17.0	.71
	LASSO	.08	.62	.24	.00	4.2	.71
	SCAD	.10	.62	.27	.03	4.3	.73
ISIS	None	.89	.97	.71	.00	17.0	.75
	LASSO	.76	.93	.22	.21	6.3	.74
	SCAD	.84	.96	.13	.49	5.6	.80
LASSO	None	.85	.97	.68	.00	15.2	.76
	LASSO	.58	.88	.29	.05	6.4	.74
	SCAD	.80	.95	.19	.19	6.0	.79
SMLE	None	.97	.99	.71	.00	17.0	.76
	LASSO	.77	.94	.21	.23	6.2	.74
	SCAD	.88	.97	.13	.45	5.7	.80

Table 2.6: Simulation results for logistic regression, setup 2

Screening Method	PLM	Coverage prop.	PSR	FDR	CSR	Ave. model size	Prediction accuracy
SIS	None	.02	.45	.89	.00	17.0	.79
	LASSO	.01	.32	.78	.00	6.1	.74
	SCAD	.01	.34	.78	.00	6.3	.75
ISIS	None	.61	.82	.81	.00	17.0	.79
	LASSO	.29	.62	.66	.00	7.4	.77
	SCAD	.56	.76	.54	.06	7.2	.81
LASSO	None	.17	.60	.84	.00	15.2	.83
	LASSO	.03	.36	.77	.00	6.4	.75
	SCAD	.09	.43	.74	.01	6.7	.76
SMLE	None	.77	.92	.78	.00	17.0	.80
	LASSO	.50	.81	.53	.01	7.1	.79
	SCAD	.76	.91	.39	.13	6.5	.83

Table 2.7: Simulation results for logistic regression, setup 3

2.5.4 Poisson regression

Parameter settings

We next consider the situation where the generic response Y follows a Poisson distribution with mean $\exp(\boldsymbol{x}^T\boldsymbol{\beta})$. Thus, data generated as independent pairs from $(Y, \boldsymbol{x})$ satisfy a Poisson regression model. The parameters used for the three correlation setups are as follows:

Setup 1: s^* is randomly chosen from $\{1, \ldots, p\}$ with $\|s^*\|_0 = 8$. For $j \in s^*$, the β_j values are generated independently from $U(\log n/\sqrt{n} + |Z|/8)$, where U is a binary random variable with $P(U = 1) = 0.8$ and $P(U = -1) = 0.2$ and Z is a standard normal random variable. For $j \notin s^*$, the β_j values are set to zero.

Setup 2: $s^* = \{1, 3, 5, 7, 9\}$. $(\beta_1, \beta_3, \beta_5, \beta_7, \beta_9) = (2, -1.8, 1.6, -1.4, 1.2)$ and $\beta_j = 0$ for $j \notin s^*$.

Setup 3: $s^* = \{1, 2, 3, 4\}$. $\beta_j = 0.7$ for $1 \leq j \leq 4$ and $\beta_j = 0$ for $j > 4$.

Similarly to the logistic example, we set $n = 200$ and $p = 1000$ for all three cases. Four screening methods (i.e., SIS, ISIS, LASSO, and SMLE) are compared with $k = [n/2 \log n]$. We used a testing likelihood ratio to measure the goodness of fit for each model. Specifically, for a given testing data set and its associated log-likelihood $\tilde{l}(.)$, we computed the testing likelihood ratio as

$$\frac{1}{T}\sum_{t=1}^{T}\left\{1 - \frac{\tilde{l}(\hat{\boldsymbol{\beta}}_t)}{\tilde{l}(\hat{\boldsymbol{\beta}}_{s^*})}\right\},$$

where $\hat{\boldsymbol{\beta}}_{s^*}$ is the MLE of $\boldsymbol{\beta}$ based on s^* and $\hat{\boldsymbol{\beta}}_t$ denotes the $\boldsymbol{\beta}$ estimate for a particular method (e.g., SMLE-SCAD) on the tth replication. We adopted the same choices for $\hat{\boldsymbol{\beta}}_t$ as in the linear example and set $T = 1000$.

Results

The results are summarized in Tables 2.8–2.10. Most of the patterns are consistent with those of the previous example. The benefits of SMLE are clear in setups 2 and 3. For setup 1, the performance of ISIS and SMLE is comparable, but SMLE is preferred because of its lower computational cost.

Screening Method	PLM	Coverage prop.	PSR	FDR	CSR	Ave. model size	Testing lh-ratio
SIS	None	.07	.75	.68	.00	19.0	.34
	LASSO	.06	.74	.30	.00	8.8	.46
	SCAD	.07	.75	.24	.02	8.1	.33
ISIS	None	.94	.99	.58	.00	19.0	.14
	LASSO	.86	.97	.20	.13	10.1	.29
	SCAD	.93	.99	.08	.52	8.7	.04
LASSO	None	.67	.94	.58	.00	18.1	.15
	LASSO	.52	.90	.32	.03	10.9	.39
	SCAD	.66	.93	.15	.29	9.0	.12
SMLE	None	.97	.99	.58	.00	19.0	.18
	LASSO	.91	.98	.14	.27	9.3	.27
	SCAD	.97	.99	.06	.63	8.5	.03

Table 2.8: Simulation results for Poisson regressions, setup 1

Screening Method	PLM	Coverage prop.	PSR	FDR	CSR	Ave. model size	Testing lh-ratio
SIS	None	.01	.53	.86	.00	19.0	.46
	LASSO	.00	.47	.64	.00	7.1	.46
	SCAD	.01	.46	.65	.00	7.1	.42
ISIS	None	.84	.95	.75	.00	19.0	.17
	LASSO	.68	.89	.41	.05	7.9	.33
	SCAD	.83	.94	.22	.29	6.3	.07
LASSO	None	.39	.82	.76	.00	17.5	.22
	LASSO	.10	.66	.57	.00	8.1	.43
	SCAD	.38	.79	.40	.06	7.0	.18
SMLE	None	.94	.98	.74	.00	19.0	.21
	LASSO	.78	.94	.33	.09	7.5	.31
	SCAD	.93	.98	.15	.37	6.0	.03

Table 2.9: Simulation results for Poisson regressions, setup 2

Screening Method	PLM	Coverage prop.	PSR	FDR	CSR	Ave. model size	Testing lh-ratio
SIS	None	.00	.19	.96	.00	19.0	.61
	LASSO	.00	.14	.93	.00	9.0	.70
	SCAD	.00	.16	.92	.00	8.2	.62
ISIS	None	.54	.69	.85	.00	19.0	.34
	LASSO	.43	.59	.69	.01	8.8	.53
	SCAD	.54	.67	.55	.09	7.4	.26
LASSO	None	.01	.26	.93	.00	17.2	.58
	LASSO	.00	.19	.91	.00	9.3	.70
	SCAD	.00	.24	.87	.00	8.1	.60
SMLE	None	.93	.98	.79	.00	19.0	.25
	LASSO	.62	.86	.56	.01	8.4	.51
	SCAD	.91	.96	.33	.14	6.2	.07

Table 2.10: Simulation results for Poisson regressions, setup 3

2.5.5 Real-data example

We now apply the SMLE-based method to a genetic application. Singh et al. [2002] measured the expression levels of 12600 genes from prostate specimens of 52 prostate cancer patients and 50 healthy controls. One objective was to build a gene-expression-based classification rule to predict the identity of unknown prostate samples. Such a classification tool is helpful in the early detection of prostate cancer, which provides a better opportunity for curative surgery. The identification of genes that influence the disease outcome also provides a greater understanding of the genetic aspect of prostate tumors.

By performing a permutation-based correlation test, Singh et al. [2002] detected 456 potential genes that are differently expressed between tumorous and normal samples. Using the 6033 genes in the complete dataset, Efron [2009] found a further 377 genes using an empirical Bayesian approach, while Chen and Chen [2012] spotted 3 more genes using EBIC-based LASSO.

In this example, we reanalyze the dataset by building a logistic regression

$$\text{logit}\{P(Y=1|\boldsymbol{x})\} = \boldsymbol{x}^T\boldsymbol{\beta},$$

where Y is the binary status of the prostate cancer (with $Y = 1$ for a tumorous sample, $Y = 0$ for a normal sample) and $\boldsymbol{x}$ contains the 12600 gene expression levels. Accordingly, we predict $Y = 1$ when $P(Y = 1|\boldsymbol{x})$ is estimated to be over 0.75 and predict $Y = 0$ otherwise. SMLE-SCAD is used for the analysis because of its superior performance in the simulation studies. We randomly select a set of 10 subjects from each of the tumorous and normal sample groups as the testing set, and treat the remainder as the training set. We set the screening bound $k = 20$

Screening Method	Ave. model size	Sensitivity	Specificity	Overall pred. error
SIS	2.2	.71	.96	.17
ISIS	2.3	.68	.96	.18
LASSO	2.7	.71	.94	.17
SMLE	2.7	.77	.94	.14

Table 2.11: Results for prostate data.

on the training set and assess the prediction accuracy of the selected model on the testing set. We base the assessment on $T = 200$ replications and summarize the results in terms of sensitivity, specificity, and the overall prediction error. For comparison purposes, we also include the results of SIS, ISIS, and LASSO followed by SCAD with $k = 20$. All the tuning parameters are selected by EBIC as in the simulation examples.

From Table 2.11, we see that all four methods performed well, choosing a parsimonious model with relatively high prediction accuracy. SMLE shows its superiority by choosing a model with higher sensitivity, which corresponds to a lower chance that a cancer patient is wrongly diagnosed as healthy. The results of SMLE are consistent with those of Chen and Chen [2012] in terms of the number of selected genes but have a lower prediction error.

2.6 Summary and conclusions

In this chapter, we have developed a new approach for the variable selection problem when the number of candidate variables is ultra-high. Our approach follows existing methods by first screening out a large number of seemingly nonsignificant covariates computationally efficiently and theoretically consistently. The screening step is followed by a widely accepted regularization approach to further reduce the number of variables in the model to enhance the interpretability. The second step results in a sequence of candidate models with increasing model complexity as the degree of regularization reduces.

Specifically, we proposed SMLE for the screening; it naturally incorporates joint effects between features in the screening process. We showed that SMLE has the sure screening property in the ultra-high-dimensional GLM setup, and we developed an iterative hard-thresholding algorithm for its implementation. Our simulation study indicated that the new procedure is computationally efficient and competitive with other screening procedures. At the same time, the new method was observed to have a higher probability of retaining all the significant covariates in the model settings that we considered.

Chapter 3

The Penalized Pseudo-Likelihood in Analysis of Survey Data

3.1 Introduction

In many areas of scientific research, one common interest is to identify the influential factors associated with certain behavioral, social, or economic indices within a target population. For example, sociologists and economists would like to identify important factors that affect the unemployment rate in a specific region, and epidemiologists are interested in finding risk behavior for diseases. In such studies, researchers often start with a survey of the target population (Korn and Graubard [1999], Rahiala and Teräsvirta [1993], Wolfson [2004]). A representative sample is then selected and measurements of the variables of interest for the sampled units are collected. A regression model is routinely employed to summarize the information contained in the data. It explains variations in the response variable through a simple function of explanatory variables (covariates). When they lack prior knowledge, researchers may collect information on many potential explanatory variables. In these applications, it is never straightforward to decide in advance which variables should be included in the perceived regression model. Consequently, the goal of identifying influential factors is often achieved through a variable selection procedure. That is, we assume a response variable together with a large number of covariates, based on which a regression model is to be fitted. In this chapter, we develop a variable selection strategy for survey data and investigate its large-sample properties.

Unlike the problems presented in Chapter 2, the sample size n is often very large, while the number of candidate covariates p is large but not huge. This makes a screening procedure unnecessary. However, the traditional all-subset selection (Section 1.3) is still computationally infeasible although its principles apply. Moreover, when we perform variable selection for survey sampling, many potential complications arise. First, the data collected through survey sampling are usually obtained from a finite population without replacement, and hence they have an intrinsic dependence structure (i.e., non-i.i.d.). Second, in complex survey designs, the inclusion probabilities of sampling units often vary across the target population. Consequently, the correlation between the response and the covariates reflected in the sample can be distorted from that of the population. This is potentially the case when some segments of the population are sampled more intensively than others. Ignoring the survey design in the selection process may result in biased selection results for the target population.

In the literature, sampling weights are often considered in estimating finite population parameters such as the population mean or population proportions. The weighted estimates help to avoid biased inference from informative sampling (Pfeffermann [1993], Fuller [2009]). However, the role of sampling weights in variable selection is not completely clear. Although the parameter estimation and variable selection serve different purposes, they often have a coherent linkage in the modeling process. It is natural to conjecture that using sampling weights is beneficial for variable selection.

In this spirit, we investigate the use of a pseudo-likelihood to take account of the sampling weights and propose a penalized pseudo-likelihood method (PPLM) for variable selection in the analysis of survey data. A sample-based BIC criterion is further derived to tune the implementation of the proposed method. In a joint randomization framework, we prove that the PPLM consistently identifies the influential variables through the BIC-based tuning. The proposed selection method is assessed through simulation studies and using data from the 2009 Survey on Living with Chronic Diseases in Canada.

This chapter is organized as follows. In Section 3.2, we introduce the joint randomization mechanism and the super-population model. In Section 3.3, we propose the PPLM and derive the sample-based BIC as a tuning strategy. In Section 3.4, we investigate the asymptotic behavior of the proposed method in a joint randomization framework. We use numerical studies in Section 3.5 to assess the performance of our approach and provide a concluding summary in Section 3.6. The proofs of the theorems are given in Appendix B, where a heuristic derivation of the proposed BIC can also be found.

3.2 Joint inference and super-population

The random behavior of an inference procedure is mostly inherited from the randomness in the data. In the context of surveys, the set of sampled units is random because of the probabilistic sampling design. At the same time, the observed value of each sampled unit may also be regarded as a random outcome from some conceptual infinite super-population (Royall [1976]).

In a design-based analysis, the finite population is regarded as nonrandom and all measurements of the sampled units are constants. The parameters of interest are finite population quantities such as the population total or the population median. The statistical inference is evaluated based on the randomness from the probability design. Nonparametric approaches are usually used for design-based inference.

One may also regard the design-induced randomness as an artifact. The measurements of the sampled units are independent realizations of a random variable from a probability model for the postulated super-population. The parameters of interest are related to the assumed model and model-based inferences are evaluated solely based on the randomization introduced by the model.

A third approach is called model-design-based inference; it incorporates randomization from both the design and the model. In such a joint randomization mechanism, the finite population is regarded as a random sample from a super-population. The survey sample is considered as a second-phase sampling from the super-population. The parameters of interest can be either model or finite-population parameters. In this mechanism, inferences on the finite-population parameters are motivated by the super-population model. Model-design-based inference can be more efficient than pure design-based approaches when the finite population is well described by the super-population model. Compared with pure model-based approaches, it protects against model violation and is therefore more robust in general (Binder and Roberts [2003], Kalton [1983]).

We study the variable selection problem under the joint randomization mechanism. Let $\mathcal{D} = \{1, \ldots, N\}$ be a finite population consisting of N sampling units. The measurements on the ith unit are denoted $(y_i, \mathbf{x}_i)$, where y_i is the response of interest and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is a p-dimensional explanatory vector (covariate vector). These are regarded as independent realizations of $(Y, \mathbf{X})$ from a super-population. We postulate a generalized linear model (GLM) on the super-population as follows. Conditioning on $\mathbf{X}$, the distribution of Y belongs to a natural exponential family, the density of which takes the form

$$f(y; \theta) = c(y) \exp\{\theta y - b(\theta)\}. \tag{3.1}$$

θ is known as the natural parameter of $f(y; \theta)$ such that $b'(\theta) = E[Y|X] \equiv \mu$ and $b''(\theta) = \text{Var}[Y|X] \equiv \sigma^2$, and $c(y)$ is a normalization constant. The influence of the explanatory variable $\mathbf{X}$ on Y is expressed through $g(\mu) = \mathbf{X}^T\boldsymbol{\beta}$ for some assumed linkage function $g(.)$, where the vector $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}^T$ is the p-dimensional regression coefficient. If $g(.)$ is the canonical link, i.e., $g(\mu) = \theta$, then we have $\theta = \mathbf{X}^T\boldsymbol{\beta}$. For simplicity, we focus on the canonical link in this chapter.

Based on this model, the effect of the explanatory variable is characterized through the size of the corresponding regression coefficient. In applications, a complex model with many variables often leads to overfitting and a poor interpretive value. Hence, it is desirable to fit the data with a parsimonious model in which many regression coefficients are estimated to be zero. Explanatory variables with nonzero coefficients are then considered to be influential on the response. To this end, we assume that $\boldsymbol{\beta}$ is ideally sparse, and address the variable selection problem by identifying a sparse model formed by the covariates with nonzero coefficients.

3.3 Pseudo-likelihood-based variable selection

3.3.1 The penalized pseudo-likelihood method

With the model settings described in Section 3.2, it is clear that, if the measurement $(y_i, \mathbf{x}_i)$ is observed for every unit in $\mathcal{D}$, the randomness in the data introduced by the probability sampling design is completely gone. In this situation, the selection of the influential variables is based on the entire population and the PLMs developed in non-survey settings (purely model-based) remain valid for the model-design-based inference. As in a typical PLM, the model coefficient vector $\boldsymbol{\beta}$ is estimated by $\check{\boldsymbol{\beta}}_\lambda$ through maximizing the penalized likelihood function

$$Q_N(\boldsymbol{\beta}) = l_N(\boldsymbol{\beta}) - N \sum_{j=1}^{p} \phi_\lambda(|\beta_j|), \tag{3.2}$$

where $l_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \log f(y_i; \mathbf{x}_i^T \boldsymbol{\beta})$ is the census log-likelihood function and $\phi_\lambda(.)$ is a penalty function indexed by a tuning parameter λ. With an appropriate choice of $\phi_\lambda(.)$ (Section 1.4.1), $\check{\boldsymbol{\beta}}_\lambda$ contains zero estimates for some coefficients and thus removes the corresponding covariates from the model.

Note that (3.2) is only conceptual, because observing $(y_i, \mathbf{x}_i)$ for all units in $\mathcal{D}$ is usually not feasible in applications. Instead, a representative sample $d = \{i_1, \ldots, i_n\} \subset \{1, \ldots, N\}$ with n units is often drawn from $\mathcal{D}$ and the measurements are observed based on the sampled units. Due to the intrinsic dependence structure among the sampled units, a full likelihood on d is prohibitive to compute in general. Alternatively, for the model-design-based inference, a pseudo-log-likelihood function is frequently used, which takes the form

$$l_n(\boldsymbol{\beta}) = \sum_{i \in d} w_i \log f(y_i; \boldsymbol{\beta}) \tag{3.3}$$

with w_i denoting the standardized survey weight for the ith unit ($\sum_{i \in d} w_i = n$). Typically, w_i is chosen proportional to $1/P(i \in d)$ such that $n^{-1} l_n(\boldsymbol{\beta})$ is design-unbiased to $N^{-1} l_N(\boldsymbol{\beta})$. Maximizing $l_n(\boldsymbol{\beta})$ over $\boldsymbol{\beta}$ leads to a maximum pseudo-likelihood estimator (MPLE) $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, i.e.,

$$\hat{\boldsymbol{\beta}} = \arg\max_{\beta} l_n(\boldsymbol{\beta}). \tag{3.4}$$

Under the appropriate sampling designs, $\hat{\boldsymbol{\beta}}$ is often $n^{-1/2}$ consistent for $\boldsymbol{\beta}$ under the joint randomization framework. The idea of using pseudo-likelihood for inference on model parameters has been widely adopted in the literature (Binder [1983], Godambe and Thompson [1986], Molina and Skinner [1992]).

Accordingly, we aim to develop an analog of PLM (3.2) based on the pseudo-likelihood. In particular, we propose the maximum penalized pseudo-likelihood estimator $\hat{\boldsymbol{\beta}}_\lambda$ that maximizes

$$Q_n(\boldsymbol{\beta}) = l_n(\boldsymbol{\beta}) - n\sum_{j=1}^{p} \phi_\lambda(|\beta_j|). \tag{3.5}$$

Compared with $Q_N(.)$, the first term in $Q_n(.)$ is the survey-weighted pseudo-likelihood, which potentially helps to avoid sampling errors that might lead to biased inferences for the target population. Meanwhile, the maximizer $\hat{\boldsymbol{\beta}}_\lambda$ of $Q_n(\boldsymbol{\beta})$ inherits the sparsity property from its census-based version $\check{\boldsymbol{\beta}}_\lambda$, which qualifies it as a variable selection operator. We refer to the selection based on $\hat{\boldsymbol{\beta}}_\lambda$ as the PPLM and further investigate its asymptotic performance in the next section.

3.3.2 Asymptotic properties of PPLM

To provide some theoretical insight into $\hat{\boldsymbol{\beta}}_\lambda$, we now establish its asymptotic consistency under the joint randomization framework described in Section 3.2. Suppose there is a sequence of finite populations, say $\mathcal{D}_r$ with $r \to \infty$. Each $\mathcal{D}_r$ is an i.i.d. sample of size N_r from a super-population modeled by (3.1) with random variable $(Y, \mathbf{X} = \{X_1, \ldots, X_p\})$. Within each $\mathcal{D}_r$, a sample d_r of size n_r is drawn according to some sampling scheme. We assume that both N_r and n_r increase to infinity as $r \to \infty$, with the sampling fraction n_r/N_r bounded by some constant $C < 1$. For simplicity of notation, we will drop the index r in the following discussion.

Without loss of generality, we assume that the first q coefficients are nonzero and denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^* = \{\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*\}$ with $\boldsymbol{\beta}_2^* = 0$. Also, we use s^* to denote the true model $\{1, \dots, q\}$ to be identified. We establish the consistency of $\hat{\boldsymbol{\beta}}_\lambda$ under the regularity conditions specified as follows, where the first two are on the super-population and the third is on the sampling plan:

C1 There exists $\xi_1 > 0$ such that

$$\max_{1 \le j \le p} E[|b''(\mathbf{X}\boldsymbol{\beta}) X_j^2|^{1+\eta}] < \infty,$$

for $\boldsymbol{\beta} \in \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*|| \le \xi_1\}$ and for some $\eta > 0$.

C2 Let

$$I(\boldsymbol{\beta}) = -E\left[\frac{\partial^2 \log f(y; \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right].$$

We assume that $I(\boldsymbol{\beta})$ is continuous at $\boldsymbol{\beta}^*$ and

$$\lambda_{min}[I(\boldsymbol{\beta}^*)] \ge M_1$$

for some constant $M_1 > 0$, where $\lambda_{min}[A]$ denotes the smallest eigenvalue of matrix A.

C3 The sampling scheme satisfies

$$\|\frac{1}{n}\sum_{i \in d} w_i z_i - \frac{1}{N}\sum_{i=1}^{N} z_i\| = O_p(n^{-\frac{1}{2}})$$

for the sequence $\{z_i\}$ such that $N^{-1}\sum_{i=1}^N |z_i|^{2+\eta} = O(1)$ with some $\eta > 0$.

Condition 2 requires that $I(\boldsymbol{\beta})$ is continuous at $\boldsymbol{\beta}^*$, so that when $\boldsymbol{\beta}$ is close enough to $\boldsymbol{\beta}^*$ the minimal eigenvalue of $I(\boldsymbol{\beta})$ is bounded away from zero. Condition 3 is quoted from Theorem 1 in Carrillo et al. [2010]; it requires that the weighted sample mean $\hat{Z}_{HT} = n^{-1}\sum_{i \in d} w_i z_i$ is root-n consistent to the population mean $\bar{Z} = \frac{1}{N}\sum_{i=1}^N z_i$. This condition is implied if asymptotic normality holds for $\hat{Z}_{HT}$, i.e., $\sqrt{n}\hat{Z}_{HT} \to_d N(\bar{Z}, \nu^2)$, which has been widely established in the literature. For example, with moment condition $N^{-1}\sum_{i=1}^N |z_i|^{2+\eta} = O(1)$, Hájek [1960] showed the asymptotic normality of $\hat{Z}_{HT}$ for simple random sampling without replacement if $n \to \infty$ and $n/N \to 0$. Bickel and Freedman [1984] established the asymptotic normality of $\hat{Z}_{HT}$ under a stratified sampling design, where samples are collected separately within strata that are prespecified in a finite population. Ohlsson [1989] further studied the behavior of $\hat{Z}_{HT}$ under a two-stage sampling framework, where primary sampling units (PSUs) are first selected from the finite population and secondary sampling units (SSUs) are collected based on the selected PSUs. Ohlsson (1989) showed that if asymptotic normality holds for $\hat{Z}_{HT}$ based on single-stage sampling on the PSUs, the asymptotic normality of $\hat{Z}_{HT}$ still holds under a corresponding two-stage sampling. For asymptotic studies of $\hat{Z}_{HT}$ under other popular sampling designs, we refer to Hájek [1964], Víšek [1979], and Chen and Rao [2007].

For the asymptotic analysis, we associate λ with n and denote the corresponding sequence by λ_n. We require the penalty function and λ_n to have the following properties:

D1 For any $\lambda > 0$, $\phi_\lambda(|\beta|) \geq 0$ for $\beta \neq 0$ and $\phi_\lambda(0) = 0$.

D2 Let $\phi'_\lambda(|\beta|) = \partial\phi_\lambda(|\beta|)/\partial|\beta|$. There exists a constant ξ_2 such that $\phi'_\lambda(|\beta|) \geq 0$ for $|\beta| \in (0, \xi_2)$ and all $\lambda > 0$. Also, $\phi'_\lambda(|\beta|)$ is continuous at β^*_j for any $j \in \{1, \ldots, q\}$.

D3 Let $\varphi_\lambda = \max\{\phi'_\lambda(|\beta_{0j}|)$ for $1 \leq j \leq q\}$. For any $M_2 > 0$, there exists ξ_3 such that
$$\phi'_{\lambda_n}(|\beta|) \geq M_2(n^{-1/2} + \varphi_{\lambda_n})$$
for $|\beta| \in (0, \xi_3)$.

With $\|.\|$ denoting the Euclidean norm, we establish the consistency of $\hat{\boldsymbol{\beta}}_{\lambda_n}$ via the following theorem.

Theorem 3.1 *Under conditions C1–C3, if $\varphi_{\lambda_n} \to 0$ as $n \to \infty$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}_{\lambda_n} = (\hat{\boldsymbol{\beta}}_{1\lambda_n}, \hat{\boldsymbol{\beta}}_{2\lambda_n})$ of the penalized pseudo-likelihood function (3.5) with $\phi_{\lambda_n}(|\beta|)$ satisfying D1–D3 such that*
$$\|\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*\| = O_p(n^{-\frac{1}{2}} + \varphi_{\lambda_n}) \quad \textit{and} \quad P\{\hat{\boldsymbol{\beta}}_{2\lambda_n} = 0\} \to 1.$$

See Appendix B for the proof. As shown in Theorem 3.1, with an appropriate choice of $\phi_\lambda(.)$, the maximum penalized pseudo-likelihood estimator is consistent in both parameter estimation and variable selection under the joint randomization framework. In particular, when $\phi_\lambda(.)$ is chosen as the L_γ penalty, i.e., $\phi_{\lambda_n}(|\beta|) = \lambda_n|\beta|^\gamma$ with $\gamma \in (0, 1)$, consistency holds if $\lambda_n \to 0$; when $\phi_\lambda(.)$ is chosen as the SCAD penalty, consistency holds if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$. In addition, for a special class of penalty functions, we have the following corollary.

Corollary 3.1 *Suppose that, for any $\beta \neq 0$, there exists $M > 0$ such that $\phi'_{\lambda_n}(|\beta|) = 0$ when $n > M$. Then, under the conditions of Theorem 3.1, the maximizer $\hat{\boldsymbol{\beta}}_{\lambda_n} = (\hat{\boldsymbol{\beta}}_{1\lambda_n}, \hat{\boldsymbol{\beta}}_{2\lambda_n})$ satisfies*

$$P(\hat{\boldsymbol{\beta}}_{1\lambda} = \hat{\boldsymbol{\beta}}_1) \to 1 \text{ and } P(\hat{\boldsymbol{\beta}}_{2\lambda_n} = 0) \to 1$$

with $\hat{\boldsymbol{\beta}}_1$ denoting the maximizer of $l_n(\boldsymbol{\beta})$ based on the true model s^*.

See Appendix B for the proof. Corollary 3.1 implies that the PPLM is able to consistently identify the influential variables and estimate their coefficients as efficiently as the MPLE (3.4) based on the true model. This result echoes the notion of the "oracle property" in Fan and Li [2001], which is desirable for the PLM in non-survey situations (Section 1.4.2).

3.3.3 Tuning strategy via sample-based BIC

As in standard PLMs, by varying the level of penalty $\phi_\lambda(.)$ in (3.5), the PPLM suggests a series of models with differing sparsity. In applications, one needs to make a choice among these sparse models. Given the specific form of $\phi_\lambda(.)$, this issue boils down to choosing an appropriate tuning parameter λ.

In the non-survey context, various criteria have been proposed for tuning a PLM (Section 1.4.3). In particular, BIC (Schwarz [1978]) has been shown to be effective (Wang et al. [2007], Zhang et al. [2010]). In the same spirit, we derive a sample-based BIC for PPLM in analysis of survey data.

Following the super-population formulation described in Section 3.2, we treat the sparse models (3.1) suggested by the PPLM as candidates for further selection. Specifically, for a specified form of penalty $\phi_\lambda(.)$, let Ω be the range of λ under consideration. We denote by s_λ a candidate model corresponding to $\hat{\boldsymbol{\beta}}_\lambda$ for some $\lambda \in \Omega$. Let $\boldsymbol{\beta}_{s_\lambda}$ be the $\tau(s_\lambda)$-dimensional coefficient of model s_λ and let ν_{s_λ} be the prior density of $\boldsymbol{\beta}_{s_\lambda}$. Then a pseudo-marginal density function of the data is given by

$$P_n(\mathbf{y}|s_\lambda) = \int L_n(\mathbf{y}; \boldsymbol{\beta}_{s_\lambda}) \nu_{s_\lambda}(\boldsymbol{\beta}_{s_\lambda}) d\boldsymbol{\beta}_{s_\lambda}.$$

Consequently, we may regard the following expression as the pseudo-posterior probability of the model s_λ:

$$P_n(s_\lambda|\mathbf{y}) = \frac{P_n(\mathbf{y}|s_\lambda)P(s_\lambda)}{\sum_{s_\lambda \in S_\Omega} P(s_\lambda)P_n(\mathbf{y}|s_\lambda)}, \tag{3.6}$$

where $S_\Omega = \{s_\lambda : \lambda \in \Omega\}$ is the collection of candidate models. In the spirit of Bayesian analysis, the model with the highest posterior $P_n(s_\lambda|\mathbf{y})$ is considered to be the one that receives the most support from the data. Since $\sum_{s_\lambda \in S_\Omega} P(s_\lambda)P_n(\mathbf{y}|s_\lambda)$ does not depend on any specific model, the highest $P_n(s_\lambda|\mathbf{y})$ is achieved by the model that maximizes the corresponding $P_n(\mathbf{y}|s_\lambda)P(s_\lambda)$. When the uniform prior $P(s_\lambda) = \zeta$ over S_Ω is used and under some regularity conditions, we obtain a Laplace approximation (Tierney et al. [1989], Kass and Raftery [1995]):

$$-2\log\{P_n(\mathbf{y}|s_\lambda)\} = -2l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) + \tau(s_\lambda)\log n + O_p(1)$$

with $\hat{\boldsymbol{\beta}}_{s_\lambda}$ denoting the MPLE (3.4) based on s_λ. Hence, we choose λ such that the corresponding model s_λ minimizes

$$\mathrm{BIC}_n(s_\lambda) = -2l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) + \tau(s_\lambda)\log n. \tag{3.7}$$

To provide further theoretical justification for the proposed BIC tuning strategy, we let s be an arbitrary model and define two sets of candidate models as follows:

- Overfitted models: $S_+ = \{s : s^* \subset s,\ s \neq s^*\}$;
- Underfitted models: $S_- = \{s : s^* \not\subset s\}$.

Then, Ω can be partitioned accordingly into

$$\Omega_+ = \{\lambda : s_\lambda \in S_+\}, \quad \Omega_- = \{\lambda : s_\lambda \in S_-\}, \quad \Omega_* = \{\lambda : s_\lambda = s^*\}. \tag{3.8}$$

In Theorem 3.1, we have shown that $P(\Omega_* \neq \varnothing) \to 1$. Therefore, selection consistency of PPLM with a tuning parameter selected by BIC (3.7) is achieved if BIC is able to identify s^* from any model s_λ with $\lambda \in \Omega_+ \cup \Omega_-$. We use the following theorem to establish this consistency result.

Theorem 3.2 *Under conditions C1–C3,*

$$P\{\min_{\lambda \in \Omega_+ \cup \Omega_-} \mathrm{BIC}_n(s_\lambda) \leq \mathrm{BIC}_n(s^*)\ \} \to 0.$$

See Appendix B for the proof.

3.4 Numerical studies

To evaluate the finite sample performance of PPLM, extensive numerical studies have been conducted using data from the Survey on Living with Chronic Diseases in Canada (SLCDC; Canada [2009]). In particular, we compare the proposed selection method with standard non-survey PLMs for a couple of sampling plans. The benefits of using survey weights are further examined in the situation where the presumed model is misspecified from the model that generates the data. We also report the analysis of the original SLCDC 2009 data as an example of using PPLM in real applications.

3.4.1 SLCDC data

SLCDC is a cross-sectional study sponsored by the Public Health Agency of Canada that collects information related to the experiences of Canadians with chronic health conditions. One of the main objectives of SLCDC is to identify health behavior that influences disease outcomes, so that the government can better plan and provide health services for people with chronic diseases.

SLCDC takes place every two years, with two chronic diseases covered in each survey cycle. The 2009 survey focused on arthritis and hypertension. We restrict our attention to hypertension. The target population for the hypertension survey is Canadians aged twenty years or older from the ten provinces who have been diagnosed with hypertension and who live in private dwellings. To facilitate the survey process, the sampling units of SLCDC 2009 are people with hypertension who completed the 2008 Canadian community health survey (CCHS). For the purpose of SLCDC, the population is first stratified according to the CCHS respondents based on sex and four age groups: 20–44, 45–64, 65–75, and 75+. Therefore, the finite population formed by the CCHS respondents was divided into 8 categories, age (4 levels) by sex (2 levels). A stratified sampling plan is used for SLCDC with proportional sample size allocation. An overall sample of 9005 was selected from the 17437 CCHS respondents, and 6142 respondents completed the SLCDC survey.

We identified 40 variables relevant to hypertension based on the original SLCDC data, of which 7 variables have complete information on all 6142 respondents. The remaining 33 variables have missing values due to non-responses in the original questionnaire (see Table 3.1 for a list of the variables and the corresponding non-response rates). There was no obvious systematic reason for the non-response. The variable with the most severe missingness is INCDRPR (household income) with a 9.6% non-response rate; the amount of missing data is relatively minor for the remaining variables. To facilitate the analysis, we used simple imputation methods for the missing data as follows. For a categorical variable, we imputed the non-response value by a random value from the response set. For a continuous variable, we imputed the non-response value by the mean value of the responses. Two exceptions to the above imputation are the variables BMHX_02 and CNHX_05. The former acts as the response variable of the regression model in the later data analysis, while the latter has natural restrictions on its range. We removed the 274 observations with missing values for these two variables, leading to basic working data with 5868 observations. The imputation/removal procedure does not have any effect on the evaluation of the BIC procedure based on the simulated population. It could bias the analysis of the real data. However, given the low rate of missingness and the plausibility of missing-at-random in this specific case, the conclusion is unlikely to be severely affected.

Since the SLCDC is a follow-up to the CCHS, the sampling weights for SLCDC were initially obtained from the weights of the CCHS data. The weights were then adjusted to ensure that the SLCDC respondents represent the target population. Consequently, the adjusted weights show considerable variation between sampled units. The standardized values of the adjusted weights vary between 0.01 and 33.62

with an inter-quartile range of 0.76.

Table 3.1: Variables for analysis of SLCDC data with non-response adjustments: A: allocate to other categories; D: delete from the data; M: impute by mean values; NA: no adjustment applied.

	Variable	Description	Levels	Missing	Adjust
1	BMHX_02	Blood pressure control status	2	1.6%	D
2	GEO_QB	Provinces grouped by region - QC	2	- -	NA
3	GEO_ON	Provinces grouped by region - ON	2	- -	NA
4	GEO_BC	Provinces grouped by region - BC	2	- -	NA
5	GEO_PR	Provinces grouped by region - PR	2	- -	NA
6	DHHX_AGE	Age	Cont.	- -	NA
7	DHHX_SEX	Sex	2	- -	NA
8	GENXDHMH	Perceived mental health	2	0.2%	A
9	CNHX_05	High blood pressure - age when diagnosed	Cont.	2.7%	D
10	MEHX_02	No. of medications taken	Cont.	0.3%	M
11	MEHX_03	No. of times per day medications taken	Cont.	0.1%	M
12	MEHXGMED	No. of medications for high blood pressure	Cont.	2.0%	M
13	MEHX_06	No. of times per day bp medication taken	Cont.	1.0%	M
14	MEHXDMCO	Medication compliance - overall	2	0.2%	A
15	HUHXDHP	Consulted family doctor about hbp	2	0.1%	A
16	SMHX_11A	Smoked at any time since being diagnosed	2	0.1%	A
17	SMHX_13A	Drank alcohol since being diagnosed	2	0.2%	A
18	SMHXDSLT	Daily salt intake	2	0.2%	A
19	SMHXDFDC	Dietary foods	2	0.1%	A
20	SMHXDPAC	Exercise/physical activity	2	0.1%	A
21	SMHXDBW	Body weight control	2	0.2%	A
22	MOHXDBPM	Self-monitoring of blood pressure	2	0.3%	A
23	MOHX_02	Correct use of bp measurement device	2	0.5%	A
24	INHX_01A	Info from family doctor	2	2.4%	A
25	INHX_01F	Info from family member/friend	2	2.4%	A
26	INHX_02A	Info from book, pamphlet, brochure	2	1.5%	A
27	INHX_02C	Info from package insert with medication	2	1.5%	A
28	INHX_02G	Info from media	2	1.5%	A
29	INHX_02H	Info from internet	2	1.5%	A
30	INHX_04	Info received - emotional impact of hbp	2	0.8%	A
31	INHX_06	Info received - correct use of medication	2	0.6%	A
32	INHX_07	Info received - additional information	2	0.9%	A
33	CPGFGAM	Gambling activity	2	0.5%	A
34	DHHDECF	Household type	2	0.2%	A
35	EDUDH04	Highest level of education in household	2	3.4%	A
36	FVCGTOT	Daily consumption - fruits and vegetables	2	5.2%	A
37	GEODUR2	Urban and rural areas	2	- -	NA
38	HWTDBMI	Body mass index (BMI) self-report	Cont.	2.1%	M
39	INCDRPR	Household income - provincial level	10	9.6%	A
40	SACDTOT	Total number hours - sedentary activities	Cont.	1.5%	M

3.4.2 Simulation settings

We first design simulation studies based on the SLCDC data. Specifically, we treat the 40 identified variables as candidate covariates for some response variable Y, and index them as X_1 to X_{40} for simplicity. We consider both continuous and binary responses in our simulations. For the continuous cases, we generate the values of Y according to

- Model 1: $Y = 0.7X_6 + 0.7X_{10} + 0.6X_{18} - 0.6X_{22} + \epsilon$,
- Model 2: $Y = 0.7X_6 + 0.6X_{10} + 0.6X_{18} - 0.5X_{22} + 0.3X_{30} - 0.3X_{34} + \epsilon$,

with $\epsilon \sim N(0, 1)$. For the binary cases where $Y \in \{0, 1\}$, we generate the values of Y according to the logistic models

- Model 3: logit(Pr$\{Y = 1|\, \mathbf{X}\}$) $= 0.7X_7 - 0.6X_8 + 0.5X_{26}$,
- Model 4: logit(Pr$\{Y = 1|\, \mathbf{X}\}$) $= 0.8X_7 - 0.7X_8 + 0.6X_{26} - 0.5X_{28} + 0.4X_{36}$.

The specified models include one of the strata identifiers in SLCDC, i.e., X_6 (Age) or X_7 (Sex). Compared with models 1 & 3, models 2 & 4 have two more influential covariates with small coefficients, so it is more difficult to correctly identify them.

The finite population used in the simulation was created as follows. The basic working data of 5868 respondents was duplicated 10 times proportional to the rounded integer values of the SLCDC weights, resulting in a pseudo-finite population of size 55950 with complete information on $X_1, \ldots, X_{40}$. The values of the response Y were then generated based on models 1–4 respectively. We consider the variable selection problem to be the identification of the postulated model that generates the values of Y.

We investigate the performance of the proposed procedure under two stratified sampling plans. Specifically, we create four strata based on variables X_6 (age, 55-/55+) and X_7 (sex, Male/Female), which leads to the group (Female, 55-) of size 7120, the group (Female, 55+) of size 19199, the group (Male, 55-) of size 6187, and the group (Male, 55+) of size 23458. In the first plan, a simple random sampling without replacement (SRSWR) with equally allocated sample sizes is drawn from each stratum. The inference is made based on the four SRSWRs pooled together. In the second plan, we further construct three subgroups within each stratum based on the sum of two binary covariates of the postulated models. The subgroups are based on $X_{18} + X_{22}$ for the data from models 1–2 and on $X_8 + X_{26}$ for the data from models 3–4. We then make inference based on the SRSWRs drawn from each subgroup of the four strata. The overall sample size is equally allocated at the stratum level with a 2:1:2 proportion for the three subgroups within the same stratum. A simple Monte Carlo computation reveals that the sample correlation between X_{18} and X_{22} (for the data from models 1–2) can be as high as 0.5, whereas their population-based correlation is around 0.02. A similar phenomenon is observed between X_8 and X_{26} (for the data from models 3–4). We therefore expect variable selection under the second sampling plan to be more challenging because of this systematic inflation. In the simulations, we set the overall sample size $n = 500$ for models 1–2 and $n = 1500$ for models 3–4.

The PPLM was then carried out on probability samples obtained from the finite population. In particular, we chose the SCAD penalty for the penalized pseudo-likelihood function (3.5), as advocated by Fan and Li [2001]. The corresponding maximization of (3.5) was solved using the thresholding-based iterative algorithm (She [2011]) and the tuning parameter was determined by the sample-based BIC (3.7). For comparison purposes, AIC (Akaike [1973]) and GCV (Craven and Wahba [1979]) were used as alternatives to the BIC tuning strategy. Based on the discussion in Section 3.3.3, we define the sample-based AIC and GCV as

$$\begin{aligned} \mathrm{AIC}_n(s_\lambda) &= -2l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) + 2\tau(s_\lambda), \\ \mathrm{GCV}_n(s_\lambda) &= -\frac{1}{n}\frac{l_n(\hat{\boldsymbol{\beta}}_{s_\lambda})}{(1-\tau(s_\lambda)/n)^2}, \end{aligned}$$

where λ was selected similarly by minimizing the corresponding scores. Moreover, for each setup, we repeated the selection procedure with all survey weights ignored (set to unity). The unweighted selection results correspond to pure model-based inferences as discussed in Section 3.2. In particular, the PPLM reduces to the standard PLM (3.2) used for non-survey situations.

3.4.3 Simulation results

In Tables 3.2–3.3, we summarize the simulation results based on 1000 repetitions in terms of the positive selection rate (PSR), false discovery rate (FDR), correct selection rate (CSR), and averaged model size (AMS). Specifically, let s_0 be the true model that generates the finite population and s'_j be the selected model based on the jth sample, $j = 1, \ldots, 1000$. The PSR, FDR, CSR, and AMS are estimated as

$$\text{PSR} = \frac{\sum_{j=1}^{1000} \tau(s^* \cap s'_j)}{1000\tau(s^*)}, \quad \text{FDR} = \frac{\sum_{j=1}^{1000} \tau(s'_j / s^*)}{1000\tau(s'_j)},$$
$$\text{CSR} = \frac{\sum_{j=1}^{1000} I(s'_j = s^*)}{1000}, \quad \text{AMS} = \frac{\sum_{j=1}^{1000} \tau(s'_j)}{1000},$$

where $\tau(s)$ denotes the size of model s and $I(.)$ is the indicator function. In addition, we assess the predictive accuracy of the selected model as follows. For each setup, a test sample of size 200 is generated by SRSWR from the same finite population as that for the training sample. For models 1–2, we use the averaged residual sum of squares (RSS) on the test data as a measurement of the predictive ability of the selected model. For models 3–4, we compute both positive and negative prediction rates. To be specific, let π^* be a specified benchmark and $\hat{\pi}_i$ be the estimated success probability of the ith test sample, $i = 1, \ldots, 200$. We then predict the ith response y_i by $\hat{y}_i = 1$ if $\hat{\pi}_i > \pi^*$ and $\hat{y}_i = 0$ otherwise. The correct prediction rates are estimated by

$$\text{PPR} = \frac{\sum_{i \in \{i: y_i = 1\}} I(\hat{y}_i = 1)}{\sum_{i=1}^{200} I(y_i = 1)}, \quad \text{NPR} = \frac{\sum_{i \in \{i: y_i = 0\}} I(\hat{y}_i = 0)}{\sum_{i=1}^{200} I(y_i = 0)}.$$

The final PPR and NPR are averaged based on 1000 replications. Note that PPR and NPR are similar to sensitivity and specificity in the clinical studies, which indicate the ability of a 0-1 prediction approach in terms of correct positive and negative predictions. In general, a larger π^* leads to high NPR but low PPR. The value of π^* should be cautiously specified in applications. In our simulation studies, we set $\pi^* = 0.5$ for simplicity.

The results are encouraging for the PPLM and the sample-based BIC tuning

Table 3.2: Selection results for the first sampling plan: Prediction assessments for models 1–2 are based on the testing RSS, while for models 3–4 they are based on (PPR, NPR) with a benchmark 0.5.

Weights	Criterion	PSR	FDR	CSR	AMS	Prediction
			Model 1			
Ignored	GCV	.96	.19	.28	4.9	1.04
	AIC	.99	.48	.05	8.7	1.08
	BIC	.96	.19	.28	4.9	1.04
Included	GCV	.95	.24	.19	5.2	1.05
	AIC	.99	.61	.01	11.4	1.11
	BIC	.95	.24	.20	5.3	1.05
			Model 2			
Ignored	GCV	.72	.19	.02	5.5	1.07
	AIC	.89	.44	.01	10.3	1.09
	BIC	.73	.19	.03	5.6	1.07
Included	GCV	.74	.24	.02	6.1	1.08
	AIC	.89	.54	.01	12.5	1.12
	BIC	.74	.24	.03	6.1	1.08
			Model 3			
Ignored	GCV	.99	.59	.00	7.8	(.71, .45)
	AIC	.99	.62	.00	8.4	(.69, .49)
	BIC	.96	.43	.00	5.1	(.72, .44)
Included	GCV	.99	.67	.00	9.9	(.71, .47)
	AIC	.99	.70	.00	10.7	(.68, .48)
	BIC	.94	.45	.00	5.3	(.71, .45)
			Model 4			
Ignored	GCV	.97	.44	.01	9.4	(.66, .55)
	AIC	.98	.47	.01	9.8	(.65, .56)
	BIC	.87	.26	.07	6.0	(.69, .53)
Included	GCV	.98	.54	.01	11.4	(.66, .54)
	AIC	.98	.56	.00	11.9	(.66, .55)
	BIC	.86	.30	.05	6.2	(.68, .53)

method. From Tables 3.2–3.3, we observe that the models selected by AIC have both high PSR and FDR, which indicates an excessive inclusion of the irrelevant variables. In comparison, BIC significantly reduces the FDR of the selected models with a slight sacrifice in the PSR, and selects models with more accurate sizes. Although GCV behaves similarly to BIC in linear model settings, it is similar to AIC

Table 3.3: Selection results for the second sampling plan: Prediction assessments for models 1–2 are based on the testing RSS, while for models 3–4 they are based on (PPR, NPR) with a benchmark 0.5.

Weights	Criterion	PSR	FDR	CSR	AMS	Prediction
			Model 1			
Ignored	GCV	.83	.23	.17	4.6	1.09
	AIC	.97	.49	.04	8.6	1.10
	BIC	.83	.23	.17	4.6	1.09
Included	GCV	.95	.31	.13	5.9	1.07
	AIC	.99	.65	.00	12.5	1.12
	BIC	.95	.30	.14	5.9	1.07
			Model 2			
Ignored	GCV	.62	.22	.02	5.0	1.13
	AIC	.88	.45	.01	10.3	1.14
	BIC	.62	.22	.02	5.1	1.12
Included	GCV	.72	.28	.01	6.5	1.10
	AIC	.89	.59	.00	13.7	1.12
	BIC	.72	.27	.01	6.5	1.10
			Model 3			
Ignored	GCV	.87	.62	.00	7.3	(.66, .44)
	AIC	.88	.63	.00	7.6	(.65, .45)
	BIC	.65	.62	.00	4.5	(.68, .42)
Included	GCV	.97	.74	.00	11.9	(.70, .46)
	AIC	.97	.75	.00	12.4	(.68, .46)
	BIC	.89	.50	.00	5.6	(.70, .44)
			Model 4			
Ignored	GCV	.94	.48	.00	9.5	(.62, .51)
	AIC	.95	.50	.00	10.0	(.62, .52)
	BIC	.72	.41	.00	6.1	(.64, .49)
Included	GCV	.93	.61	.00	12.5	(.64, .53)
	AIC	.94	.62	.00	12.9	(.64, .53)
	BIC	.82	.34	.01	6.4	(.67, .54)

for the logistic models where less information is provided by the binary responses.

In the first sampling plan, the inclusion probabilities are related to Y through a single covariate in the model (i.e., X_6 or X_7). The sample correlation structure between the response and covariates is largely maintained from the finite population. Consequently, no substantial difference is observed between the weighted

Table 3.4: Selection frequency of influential variables in model mis-specified case

Weights	Criterion	X_{18}	X_{20}	X_{38}	AMS	Testing RSS
			n=500			
Ignored	GCV	.78	.95	.56	5.9	1.93
	AIC	.95	.99	.73	12.5	1.95
	BIC	.83	.97	.60	6.6	1.93
Included	GCV	.73	.92	.84	6.3	1.77
	AIC	.91	.99	.85	12.5	1.79
	BIC	.78	.94	.83	6.9	1.77
			n=1000			
Ignored	GCV	.96	1.00	.79	7.6	1.87
	AIC	.99	1.00	.87	13.1	1.88
	BIC	.97	1.00	.80	7.9	1.87
Included	GCV	.93	1.00	.94	7.6	1.71
	AIC	.98	1.00	.96	13.0	1.72
	BIC	.94	1.00	.94	7.7	1.71

and unweighted selection procedures from Table 3.2.

The benefits of using sampling weights in variable selection are tentatively revealed by the second sampling plan, where the sample correlation structure is systemically distorted. Clearly, the spurious correlation between covariates in the sampled units deteriorates the efficiency of selection methods. This is reflected by the reduced PSRs and the inflated FDRs for the unweighted procedures. Incorporating sampling weights in the selection process helps to correct the biased result. In particular, noticeable improvements have been observed for the PPLM. In the most impressive case (i.e., model 3 of Table 3.3), the PPLM with BIC substantially improves the standard PLM procedure by increasing the PSR from .65 to .89, while reducing the corresponding FDR from .62 to .50. Our observation echoes the rationale for weighting to remove bias due to informative sampling Fuller [2009].

Sampling weights also provide protection against model mis-specification (Pfeffermann and Holmes [1985], Kott [1991]): inferences based on weighted estimates may remain valid for the surveyed population, even when the model fails. To gain further insight into weighting in variable selection, we compare the PPLM with the standard PLM in a simulation where the presumed model is misspecified from the model that generates the data. In this situation, a postulated "true" model does not exist, and the goal of variable selection is to find an optimal model that well describes the finite population. We still make use of the stratified pseudo-finite population in Section 3.4.2 but generate the response variable Y according to the strata. Specifically, the values of Y for units in strata (Male, 55+) and (Female, 55+) were generated by

$$Y = 0.6X_6 + 0.4X_{18} + 0.4X_{20} + 0.6X_{38} + \epsilon,$$

while the values Y for units in the strata (Male, 55-) and (Female, 55-) were generated by

$$Y = 0.6X_6 + 0.4X_{18} + 0.4X_{20} + \epsilon$$

with $\epsilon \sim N(0, 1)$ denoting a random error. In other words, we assume that variable X_{38} is influential for people aged 55 and older but not for those younger than 55. In addition, we further violate the presumed model 3.1 by excluding X_6 from the set of candidate covariates, which mimics the situation where one important design feature is omitted in the modeling. A stratified SRSWR of size 500 or 1000 is drawn using the first sampling plan in Section 5.2. The weighted and unweighted procedures are then tested for the variable selection based on the sampled units.

We summarize the simulation results in Table 3.4 by estimating the selection rates of X_{18}, X_{20}, and X_{38} based on 1000 replications. Similarly to the previous simulations, the averaged model size (AMS) and the testing RSS of the selected models (i.e., the averaged RSS based on testing data of size 200) are also included in the summary. From Table 3.4, we see that when the model assumption is violated, the PPLM still achieves relatively high prediction accuracy by suggesting relevant variables with high probability. In contrast, ignoring the survey weights leads to a relative loss of nearly 9% on the testing RSS because of the exclusion of X_{38}. Apparently, increasing the sample size helps to improve the goodness of fit for the misspecified models, but the cost is the inclusion of more variables.

3.4.4 Analysis of SLCDC data

To illustrate the application of PPLM, we use it to identify health behaviors that affect the control of blood pressure using SLCDC 2009. The response variable is BMHX_02 from the working data obtained from SLCDC, which has two levels indicating whether or not the blood pressure of the respondent is under control, based on the latest measurement by a health professional. We treat the remaining 39 variables in the working data as candidate covariates, and our goal is to identify the influential covariates that are associated with blood-pressure control. We build a logistic regression of BMHX_02 on the candidate covariates and use PPLM with the SCAD penalty to select the influential covariates. As a preliminary step, each covariate is standardized such that the corresponding first and second weighted sample moments are zero and unity respectively. For comparison, the BIC (3.7) and AIC/GCV are used for the tuning parameter selection.

In Figure 3.1, we plot the criterion scores with respect to the model sparsity. We see that BIC selects a model with 11 covariates, while GCV and AIC select the same model, which has 24 covariates. When survey weights are ignored in the selection procedure, models with 7 or 21 covariates are suggested based on PLM with the standard BIC or GCV/AIC. The difference between the weighted and unweighted selection results reflects the potential distortion in the correlation structure of the sampled units. This difference may also be explained by the model mis-specification for part of the SLCDC population (Lohr and Liu [1994]). Given the potential bias of unweighted methods, the weighted results are more plausible.

We further assess the selected models in terms of their predictive accuracy as follows. First, we draw 500 independent sets of 5868 bootstrap samples (with replacement) from the working data of SLCDC. For the tth bootstrap sample d_t, $t = 1, \dots, 500$, the survey weight w_i for the ith unit is adjusted by $\tilde{w}_{ti} = v_{ti} w_i$ with v_{ti} denoting the number of times that the ith unit is selected in d_t. We then fit the selected models to each bootstrap sample (with the weights accounted for accordingly), and evaluate their weighted positive and negative prediction rates (WPPR, WNPR) by

$$\text{WPPR} = \frac{\sum_{i \notin d_t} w_i I(\hat{y}_i = 1, y_i = 1)}{\sum_{i \notin d_t} w_i I(y_i = 1)}, \quad \text{WNPR} = \frac{\sum_{i \notin d_t} w_i I(\hat{y}_i = 0, y_i = 0)}{\sum_{i \notin d_t} w_i I(y_i = 0)},$$

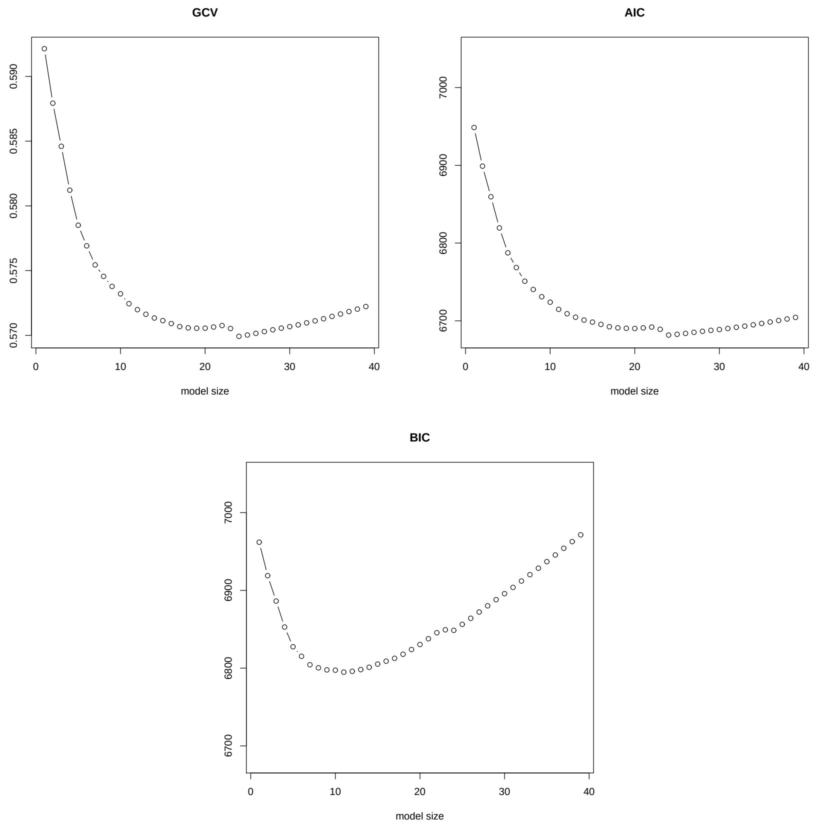

Figure 3.1: Selection criteria values based on candidate models

where y_i and $\hat{y}_i$ denote the ith response in BMHX_02 and its predicted value. We summarize the averaged WPPR and WNPR based on 500 bootstrap samples in Table 3.5 according to three different benchmark values (i.e., 0.25, 0.35, 0.45).

From Table 3.5, we observe that the models selected from the unweighted analysis have a lower WPPR in general, which provides additional support for using survey weights in the selection procedure. Compared with GCV/AIC, BIC selects a model with a slightly conservative WPPR but a higher WNPR. However, the difference is not significant. The size of the BIC-selected model is much less than that of the model selected by GCV/AIC, which provides an easier interpretation of the response BMHX_02 and the covariates.

To assess the stability of the selection, we repeat the PPLM based on 500 boot-

Table 3.5: Prediction accuracy for selected models (WPPR, WNPR) based on different benchmarks.

Weights	Criteria	$\geqslant .25$	$\geqslant .35$	$\geqslant .45$
Ignored	AIC/GCV	(.646, .525)	(.460, .688)	(.299, .811)
	BIC	(.649, .513)	(.445, .705)	(.265, .818)
Included	AIC/GCV	(.645, .523)	(.488, .682)	(.338, .790)
	BIC	(.654, .532)	(.485, .706)	(.322, .830)

Table 3.6: Bootstrap selection results for significant variables: (Estimated coefficient, Standard error, Selection rate).

Variable	GCV	AIC	BIC
GEO_ON	(.14, .09, .86)	(.16, .09, .92)	(.09, .09, .58)
DHHX_AGE	(-.29, .09, 1.0)	(-.32, .09, 1.0)	(-.27, .08, 1.0)
GENXDHMH	(-.15, .05, .99)	(-.15, .05, .99)	(-.14, .06, .92)
SMHXDSLT	(.11, .07, .76)	(.12, .07, .84)	(.08, .09, .47)
MOHXDBPM	(-.08, .07, .67)	(-.09, .06, .81)	(-.05, .07, .35)
INHX_06	(.18, .06, .97)	(.18, .06, .99)	(.18, .07, .91)
HWTDBMI	(.14, .06, .95)	(.14, .06, .97)	(.13, .06, .91)
Ave. Model Size	23.1	27.8	10.3

strap samples. In Table 3.6, we list the bootstrap selection rate for the seven most significant covariates according to their MLEs in the original SLCDC working data. The corresponding coefficient estimates and standard errors are also included based on the bootstrap samples. From Table 3.6, we find that only four significant covariates (i.e., DHHX_AGE, GENXDHMH, INHX_06, and HWTDBMI) are consistently selected by BIC, while GCV/AIC tends to select more unreliable covariates. The selection results based on BIC suggest a strong association of blood pressure control with age, weight, mental health, and medication information; this is consistent with the hypertension literature (Gelber et al. [2007], Yan et al. [2003]).

3.5 Summary and conclusions

In this chapter, we have addressed variable selection in the analysis of complex surveys. When units are selected by disproportionate sampling, the data correlation structure in the sample can be distorted. Incorporating sampling weights in the selection process protects against biased selection results. In this spirit, we proposed a survey-weighted regularization method based on the pseudo-likelihood (PPLM) and derived a sample-based BIC for its implementation. Under some regularity conditions, we showed that the PPLM consistently identifies the influential variables under a joint randomization framework. The performance of the proposed method was confirmed by numerical studies.

Chapter 4

A Thresholding Algorithm for PLM in Finite Mixture Models

4.1 Introduction

In the previous chapters, we have addressed the feature selection problem for ultra-high-dimensional data and complex survey data. New approaches have been developed in the framework of penalized likelihood methods (PLMs), and their effectiveness has been illustrated via both theoretical and numerical studies. In this chapter, we continue our investigation of PLMs to address feature selection for finite mixture models in the analysis of heterogeneous data.

Finite mixture models play important roles in statistical learning and data mining. They have important applications in scientific disciplines such as genetics, marketing, and medical research (Shoukri and MacLachlan [1994], Böhning [2000], Brijs et al. [2004]). They are routinely used to detect the presence of subpopulations in an overall population. Determining the number of components (order selection) is a fundamental problem in these applications. For instance, the order of a mixture model in a genetic application is indicative of some fundamental gene structures. In financial applications, the order reflects the complexity needed to appropriately model the real world and thereby guides the relevant practices. A mixture model with an excessive number of components usually overfits the data and leads to poor interpretive value. Many researchers have investigated the strategies that determine the appropriate order of a finite mixture model given a random sample from the target population.

Order selection is invariably a trade-off between model complexity and goodness-of-fit. The classical Akaike information criterion (AIC; Akaike [1973]) and the Bayesian information criterion (BIC; Schwarz [1978]) are often employed in the context of finite mixture models. These methods discount the likelihood-based measure of the goodness-of-fit by penalties directly proportional to the model order. Although BIC has been shown to be consistent (Leroux [1992], Keribin [2000]), its optimality under regular models does not extend to nonregular finite mixture models for order selection under general conditions. Many new procedures have been discussed in the literature, such as distance-measure-based approaches (Chen and Kalbfleisch [1996], Woo and Sriram [2006]), hypothesis-testing-based approaches (Ghosh and Sen [1985], MacLachlan [1987], Chen and Chen [2001], Li and Chen [2010]) and Bayesian approaches (Richardson and Green [1997], Ishwaran et al. [2001], Berkhof et al. [2003]). In this chapter, we study the penalized likelihood method (PLM) introduced by Chen and Khalili [2008]. Unlike other approaches, this method introduces penalties based on two types of overfitting. In particular, it introduces a nonsmooth penalty term to merge close subpopulations in a finite mixture model. The shrinkage principle of regularized regression, e.g., LASSO (Tibshirani [1996]) and SCAD (Fan and Li [2001]), is seamlessly employed for the order selection. With an appropriate choice of the penalty functions, the method is consistent in both identifying the order of the mixture model and estimating the mixing distribution. It has the advantage that the order selection and parameter estimation are completed in one strike.

Similarly to PLMs in the context of regression, the numerical problem of Chen and Khalili [2008] does not have straightforward solutions because the penalty functions can be nonsmooth and nonconvex. In the case of SCAD for regression models, Fan and Li [2001] propose locally approximating the nonconvex penalty by a quadratic function. With the aid of local quadratic approximation (LQA), they solve a series of convex optimization problems over smooth objective functions. Chen and Khalili [2008] adopt this strategy and find that the method is also suitable for order selection. However, it does not directly give sparse solutions, as required for variable or order selection. Recently, substantial progress has been made on optimization problems related to regularization methods (Efron et al. [2004], Zou and Li [2008], Friedman et al. [2007]). In particular, thresholding-based algorithms (Daubechies et al. [2004], She [2009]) provide a superior solution to the order selection problem.

In this chapter, we aim to provide an effective computational approach for the regularization-based order selection method in finite mixture models. The implementation of such methods can be embedded in a typical EM framework, and the challenge arises within the M-step, where the objective function is multivariate, nonsmooth, and nonconcave. To overcome this difficulty, we first transform the multivariate optimization problem into a set of univariate optimizations. A thresholding-based algorithm is then used for the nonsmooth and nonconcave but univariate objective functions. We refer to this new computational strategy as the iterative thresholding-based descent algorithm (ITD). Within the EM-framework, the ITD efficiently leads to a sparse estimate of the mixing distribution. It hence attains the goal of the order selection of Chen and Khalili [2008]. We establish the convergence of ITD and demonstrate its performance through simulations and real-data examples.

The rest of this chapter is organized as follows. In Section 4.2, we review the regularization approach for order selection and discuss implementation strategies. We introduce the ITD in Section 4.3 and establish its convergence. We discuss ITD tuning issues in Section 4.4. In Section 4.5, we illustrate the ITD via simulations and examples, and concluding remarks are given in Section 4.6. The proofs of the theorems are given in Appendix C.

4.2 Order selection via regularization

4.2.1 Mixture model and penalized likelihood

A finite mixture model of order K is a concave combination of K standard probability density functions. We focus on the situation where the component distribution is from an exponential family with a possible dispersion parameter:

$$g(y;\theta,\phi) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)] \tag{4.1}$$

with respect to some σ-finite measure and specific functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$. Let Θ and Φ be the parameter spaces for θ and ϕ. We consider the case $\Theta \subset R$ and $\Phi \subset R^{+}$, which includes normal, Poisson, binomial, and many other widely used distribution families.

The density function of a finite mixture model with order K is

$$f(y; \boldsymbol{\theta}, \boldsymbol{\pi}, \phi) = \sum_{k=1}^{K} \pi_k g(y; \theta_k, \phi), \tag{4.2}$$

where $g(y; \theta_k, \phi)$ specifies the kth component density function with component parameter θ_k and shared structure parameter ϕ. We use $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ for the mixing proportions, with $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\theta}$ for the vector of component parameters. The corresponding mixing distribution on Θ assigns probability π_k to value θ_k. We assume that $\pi_k \neq 0$ and $\theta_j \neq \theta_k$ for all $j \neq k$. For simplicity of notation, we denote $\boldsymbol{\Psi} = (\boldsymbol{\theta}, \boldsymbol{\pi}, \phi)$.

Given a set of i.i.d. observations $Y = (y_1, \ldots, y_n)$ from (4.2), the log-likelihood of the mixing distribution with order K is given by

$$l_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}, \boldsymbol{\pi}, \phi). \tag{4.3}$$

Maximizing $l_n(\boldsymbol{\Psi})$ over $\boldsymbol{\Psi}$ leads to a nonparametric maximum likelihood estimator (MLE) with a finite order (Lindsay [1983], Lesperance and Kalbfleisch [1992]). However, such an MLE provides little information on the actual order of the mixture model. It may overfit by assigning a negligible proportion to an arbitrary subpopulation (Type 1) and/or by including several nearly identical subpopulations in the model (Type 2).

Chen and Khalili [2008] introduce penalties to prevent both types of overfitting. In particular, they introduce a regularization penalty to merge close subpopulations to reduce the model complexity. To be specific, for any prespecified large K, we denote the component parameters $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_K$. Let $\eta_k = \theta_{k+1} - \theta_k$ for $k = 1, \ldots, K-1$. Chen and Khalili [2008] propose a regularization method that maximizes the penalized log-likelihood function

$$\tilde{l}_n(\boldsymbol{\Psi}) = l_n(\boldsymbol{\Psi}) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_\lambda(|\eta_k|), \tag{4.4}$$

where $C_K > 0$ is a scale constant and $p_\lambda(\cdot)$ is a nonsmooth penalty function with a spike at 0. As in regularized regression analysis, when $p_\lambda(\cdot)$ spikes at 0, the penalized log-likelihood $\tilde{l}_n(\boldsymbol{\Psi})$ has a positive probability of attaining its maximum with some $\eta_k = 0$. Hence, a built-in order selection procedure is obtained by discouraging Type 2 overfitting. By tuning the level of the penalty λ in $p_\lambda(.)$, we arrive at a finite mixture model with a suitable order together with a maximum penalized likelihood estimator (MPLE).

The first penalty in (4.4) prevents Type-1 overfitting by forcing the mixing proportions away from zero [Chen and Kalbfleisch, 1996]. Consequently, the resulting model has θ_k clustered around the parameters of the true subpopulations. This leads to some small values of η_k to be squeezed by the second penalty $p_\lambda(\cdot)$. A finite mixture model with a lower order is hence attained.

4.2.2 The EM algorithm

The implementation of the maximization of $\tilde{l}_n(\boldsymbol{\Psi})$ can be embedded in a typical EM framework (Dempster et al. [1977]).

Let z_{ik} for $i = 1, \ldots, n$, $k = 1, \ldots, K$ be a 0-1 variable showing the component membership of the ith observation. If all the z_{ik} are observed together with y_i, the log-likelihood of the complete data is given by

$$l_n^c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log \pi_k + \log\{g(y_i; \theta_k, \phi)\}\right].$$

The penalized log-likelihood of the complete data takes the form

$$\tilde{l}_n^c(\boldsymbol{\Psi}) = l_n^c(\boldsymbol{\Psi}) + C_K \sum_{k=1}^{K} \log \pi_k - \sum_{k=1}^{K-1} p_\lambda(|\eta_k|).$$

With an initial value of $\boldsymbol{\Psi}^{(0)}$, the EM algorithm maximizes $\tilde{l}_n(\boldsymbol{\Psi})$ through the following steps based on $\tilde{l}_n^c(\mathbf{\Psi})$:

- *E-step*: Let $\mathbf{\Psi}^{(t)}$ be the estimate of the parameters after t iterations. Compute $Q(\mathbf{\Psi};\mathbf{\Psi}^{(t)})$, the conditional expectation of $\tilde{l}_n^c(\mathbf{\Psi})$ given Y with $\mathbf{\Psi}^{(t)}$ as the true value of the model parameters:

$$\begin{aligned} Q(\mathbf{\Psi};\mathbf{\Psi}^{(t)}) &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)} \log\{g(y_i;\theta_k,\phi)\} \\ &\quad + \sum_{i=1}^{n}\sum_{k=1}^{K}\{w_{ik}^{(t)} + \frac{C_K}{n}\}\log \pi_k - \sum_{k=1}^{K-1} p_\lambda(|\eta_k|), \end{aligned}$$

where

$$w_{ik}^{(t)} = E(z_{ik}|Y) = \frac{\pi_k^{(t)} g(y_i;\theta_k^{(t)},\phi^{(t)})}{\sum_{l=1}^{K}\pi_l^{(t)} g(y_i;\theta_l^{(t)},\phi^{(t)})}.$$

- *M-step*: The M-step maximizes $Q(\mathbf{\Psi};\mathbf{\Psi}^{(t)})$ over $\mathbf{\Psi}$. We find

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(t)} + C_K}{n + KC_K}$$

and

$$\phi^{(t+1)} = \arg\max_{\phi}\left\{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)}\log\{g(y_i;\theta_k^{(t)},\phi)\}\right\}, \tag{4.5}$$

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}}\left\{\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(t)}\log\{g(y_i;\theta_k,\phi^{(t)})\} - \sum_{k=1}^{K} p_\lambda(|\eta_k|)\right\} \tag{4.6}$$

The above EM procedure leads to a local maximizer of $\tilde{l}_n(\mathbf{\Psi})$ by iteratively repeating the E-step and M-step. See Wu [1983] and MacLachlan and Krishnan [2008] for the convergence of EM-based computational procedures.

The EM-algorithm appears to have completely addressed the numerical issue, but it contains a weakness (4.6). The optimization problem (4.6) is similar to the numerical problem in PLMs for regression analysis. Naturally, Chen and Khalili [2008] employ the local quadratic approximation (LQA) method, which is designed primarily for SCAD. However, LQA does not directly provide a sparse solution for the variable selection or a finite mixture model with an appropriate order. An additional step is thus required to manually set near-zero values to exact zero. The development of an effective and direct numerical algorithm is the topic of this chapter.

4.3 Iterative thresholding descent procedure

As reviewed in Section 1.4.4, substantial progress has been made in solving the optimization problems related to PLMs. The stagewise least angular regression (LARS; Efron et al. [2004]) is the first breakthrough for LASSO. Zou and Li [2008] suggest a weighted L_1 approximation for SCAD, such that the corresponding nonconcave maximization can be carried out by the efficient algorithms designed for LASSO. Most recently, the coordinate-descent-based methods (Friedman et al. [2007]) have dramatically sped up the computation for LASSO, and they work even when the number of variables far exceeds the sample size in the regression analysis.

Although these methods are not appropriate for our numerical problem, they are a source of inspiration. In particular, we find that the thresholding-based method (Daubechies et al. [2004], She [2009]) can be engineered to address order selection in finite mixture models. This algorithm has two important steps. The first step is to translate the multivariate objective function into an equivalent auxiliary function that can be decomposed into a sum of several univariate functions. This step dramatically simplifies the problem and reduces the computation. The second step is a thresholding procedure that iteratively performs simple thresholding operations on these univariate functions. This is useful particularly when the objective function is nonsmooth and nonconcave. When the procedure converges, a sparse solution is obtained and the goal of order selection is achieved.

We now return to the core issue (4.6), which can be rewritten as

$$\min_{\boldsymbol{\theta}} \left\{ Q(\boldsymbol{\theta}) = -\sum_{k=1}^{K} \varphi_k(\theta_k) + a(\phi) \sum_{k=1}^{K-1} p_\lambda(|\theta_{k+1} - \theta_k|) \right\}, \tag{4.7}$$

with

$$\varphi_k(\theta_k) = \sum_{i=1}^{n} w_{ik}\{y_i\theta_k - b(\theta_k)].$$

It can easily be seen that $Q(\boldsymbol{\theta})$ is a multivariate function containing a nonsmooth and usually nonconvex penalty function $p_\lambda(\cdot)$. Our first step toward an effective algorithm is to create an auxiliary function that is a sum of univariate functions.

Let $\eta_0 = \theta_1$, $\eta_k = \theta_{k+1} - \theta_k$, so that $\theta_k = \sum_{j=0}^{k-1} \eta_j$ for $k = 1, \ldots, K$. After this parameter transformation, the objective function in (4.7) becomes

$$Q(\boldsymbol{\eta}) = -\sum_{k=1}^{K} \varphi_k \left(\sum_{j=0}^{k-1} \eta_j \right) + a(\phi) \sum_{j=1}^{K-1} p_\lambda(\eta_j). \tag{4.8}$$

Next, setting $\zeta_k = \sum_{j=0}^{k-1} \xi_j$ for $k = 1, \ldots, K$, we introduce an auxiliary function in $\boldsymbol{\xi} = (\xi_0, \ldots, \xi_{K-1})$,

$$\begin{aligned} G(\boldsymbol{\xi}; \boldsymbol{\eta}) \quad &= \quad uQ(\boldsymbol{\xi}) + \frac{1}{2} \sum_{j=0}^{K-1} (\xi_j - \eta_j)^2 \\ &\quad - u \sum_{k=1}^{K} \sum_{i=1}^{n} w_{ik}\{b(\zeta_k) - b(\theta_k) - b'(\theta_k)(\zeta_k - \theta_k)\} \qquad (4.9) \end{aligned}$$

for some positive scale u. The auxiliary function has two crucial properties. First, we notice that $Q(\boldsymbol{\eta}) = u^{-1}G(\boldsymbol{\eta}; \boldsymbol{\eta})$. This property links a stationary point defined through G to a local minimum of Q. Second, because of the specific form of $Q(\boldsymbol{\xi})$, it can be seen that $G(\boldsymbol{\xi}; \boldsymbol{\eta})$ is additive in the components of $\boldsymbol{\xi}$. This property makes it simple to maximize $G(\boldsymbol{\xi}; \boldsymbol{\eta})$ with respect to $\boldsymbol{\xi}$ for any given $\boldsymbol{\eta}$ through a thresholding procedure.

We now make the first point. Let $\boldsymbol{\eta}^{(0)}$ be an initial vector of $\boldsymbol{\eta}$. For $m = 1, 2, \ldots$, define

$$\boldsymbol{\eta}^{(m)} = \arg\min_{\boldsymbol{\xi}} G(\boldsymbol{\xi}; \boldsymbol{\eta}^{(m-1)}). \tag{4.10}$$

The additivity of $G(\boldsymbol{\xi}; \boldsymbol{\eta})$ with respect to $\boldsymbol{\xi}$ decomposes the above operation into a set of univariate optimization problems:

$$\begin{cases} \min_{\xi_0} \ \left(\xi_0 - \left[\eta_0 + u\sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik}[y_i - b'(\theta_k)]\right]\right)^2, \\ \min_{\xi_j} \frac{1}{2}\left(\xi_j - \left[\eta_j + u\sum_{k=j+1}^{K}\sum_{i=1}^{n} w_{ik}[y_i - b'(\theta_k)]\right]\right)^2 + ua(\phi)p_\lambda(|\xi_j|) \end{cases} \tag{4.11}$$

for $j = 1, \dots, K-1$. Clearly, these optimization problems have a unified form

$$\min_{\gamma} \left\{ q(\gamma) = (\gamma - z)^2 + \kappa p_\lambda(|\gamma|) \right\} \tag{4.12}$$

with $\kappa = 0$ for the case $j = 0$, and $\kappa = ua(\phi)$ otherwise. A threshold technique will be used to overcome the difficulty caused by the nonsmooth and nonconvex penalty function $p_\lambda(\cdot)$.

Recall that $\boldsymbol{\eta}$ can be linearly transformed to $\boldsymbol{\theta}$ by $\boldsymbol{\theta} = \Gamma\boldsymbol{\eta}$, where Γ is a lower triangular matrix with the elements below the diagonal equal to one. The following theorem shows that the iteration defined in (4.10) reduces the value of Q.

Theorem 4.1 *Following the settings and notation introduced earlier, assume that* $b(.)$ *in (4.1) is twice continuously differentiable over* Θ. *Let* $\boldsymbol{\eta}^{(m)}$ *be the sequence defined by (4.10) and* $\boldsymbol{\theta}^{(m)} = \Gamma\boldsymbol{\eta}^{(m)}$. *Denote by* τ_1 *the maximum eigenvalue of* $\Gamma^T\Gamma$, *and let*

$$\tau_2^{(m)} = \max_k \sup_{0<\alpha<1} b''(\alpha\theta_k^{(m+1)} + (1-\alpha)\theta_k^{(m)}).$$

If $0 < u < [n\tau_1\tau_2^{(m)}]^{-1}$, *then*

$$Q(\boldsymbol{\eta}^{(m+1)}) \leq Q(\boldsymbol{\eta}^{(m)}),$$

and equality holds only when $\boldsymbol{\eta}^{(m+1)} = \boldsymbol{\eta}^{(m)}$.

See Appendix C for the proof. If $Q(\boldsymbol{\eta})$ is bounded from below, then $Q(\boldsymbol{\eta}^{(m)})$ is a monotone decreasing sequence that must therefore converge to a limit. If in addition, $\{\boldsymbol{\eta} : Q(\boldsymbol{\eta}) < Q(\eta^{(0)})\}$ is compact, then $\{\boldsymbol{\eta}^{(m)}\}$ must have a convergent subsequence. Many GLMs lead to a regularized likelihood function Q satisfying these two conditions. Thus, this is likely to provide an optimization solution. We are interested further in whether $\{\boldsymbol{\eta}^{(m)}\}$ itself converges and whether the limit of $Q(\boldsymbol{\eta}^{(m)})$ is a local minimum of $Q(\cdot)$. The following corollary confirms the first point.

Corollary 4.1 *Assume the conditions of Theorem 4.1 and that* $\{\boldsymbol{\eta} : Q(\boldsymbol{\eta}) \leq Q(\boldsymbol{\eta}^{(0)})\}$ *is compact. Let* $\tau_2^* = \sup_m(\tau_2^{(m)})$*. When* $u \in (0, [n\tau_1\tau_2^*]^{-1})$*, the sequence* $\{\boldsymbol{\eta}^{(m)}\}$ *is asymptotically regular. That is, as* $m \to \infty$*,*

$$\|\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}\| \to 0.$$

See Appendix C for the proof. Corollary 4.1 not only serves as a necessary step toward a complete proof of the convergence of $\{\boldsymbol{\eta}^{(m)}\}$, but also provides a straightforward stopping criterion for procedure (4.10). That is, we can terminate if $\|\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}\| < \varepsilon$ for some prespecified tolerance level ε. We set $\varepsilon = 10^{-5}$ in our implementation.

Using Corollary 4.1, we now prove the convergence of the ITD procedure via the following theorem.

Theorem 4.2 *Under the conditions of Corollary 4.1, and if the number of stationary points of the objective function* $Q(\boldsymbol{\eta})$ *is finite, then the sequence* $\{\boldsymbol{\eta}^{(m)}\}$ *converges to a fixed point of (4.10) that is also a stationary point of* $Q(\boldsymbol{\eta})$*.*

See Appendix C for the proof. Theorems 4.1 and 4.2 are more general than the specific iteration scheme (4.10). Two assumptions, that $\{\boldsymbol{\eta} : Q(\boldsymbol{\eta}) \leq Q(\boldsymbol{\eta}^{(0)})\}$ is compact and that the number of stationary points of $Q(\cdot)$ is finite, are trivially satisfied for exponential families. However, unless $Q(\cdot)$ is a convex function, there is no guarantee that $\boldsymbol{\eta}^{(m)}$ converges to a global minimum. Multiple initial values are often used in the hope that the global minimum will be found.

The above results are useful only if we carry out (4.10) effectively. This task is made simpler because $G(\boldsymbol{\xi}; \boldsymbol{\eta})$ is additive in the elements of $\boldsymbol{\xi}$. We need only to solve the standard optimization problem (4.12).

There is an intrinsic link between the solution to (4.12) and the thresholding rule in wavelet applications (Antoniadis [2007]). Let z_+ be the positive part and sgn(z) be the sign of z for any real number z. With $\kappa = 1$, when the L_1 penalty

$p_\lambda(\gamma) = \lambda|\gamma|$ is used, the solution of (4.12) is given by the soft thresholding rule as $\gamma^* = (|z| - \lambda)_+ \cdot \mathrm{sgn}(z)$; and when the L_0 penalty $p_\lambda(\gamma) = \lambda^2/2 \cdot I(\gamma \neq 0)$ is used, the solution of (4.12) is given by the hard thresholding rule as $\gamma^* = z \cdot I(|z| > \lambda)$. Explicit solutions of (4.12) for other commonly used penalty functions are available in the literature (Antoniadis [2007], Antoniadis and Fan [2001]). Since the SCAD penalty was advocated by Chen and Khalili [2008] for their regularization method (namely, MSCAD), we give the corresponding explicit solution as follows.

The SCAD penalty is defined as a symmetric function with $p_\lambda(0) = 0$ and for $\gamma > 0$,

$$p_\lambda^{'}(\gamma) = \lambda I(\gamma \leq \lambda) + \frac{(\nu\lambda - \gamma)_+}{(\nu - 1)} I(\gamma > \lambda) \tag{4.13}$$

for some constant $\nu > 2$ and a tuning parameter λ. The exact size of ν does not have a noticeable effect on the performance of the order selection, while the level of λ needs to be further specified in applications (see Section 4.2 for a discussion).

Proposition 4.1 *When $0 < \kappa \leq \nu - 1$, the solution to (4.12) is given by*

$$\gamma^* = \begin{cases} (|z| - \kappa\lambda)_+ \cdot sgn(z), & when\ |z| < (\kappa+1)\lambda \\ \frac{(\nu-1)z - \kappa\nu\lambda sgn(z)}{\nu-\kappa-1}, & when\ (\kappa+1)\lambda \leq |z| < \nu\lambda \\ z, \quad when\ |z| \geq \nu\lambda. \end{cases}$$

When $\nu - 1 < \kappa \leq \nu$, the solution to (4.12) is given by

$$\gamma^* = \begin{cases} (|z| - \kappa\lambda)_+ \cdot sgn(z), & when\ |z| < \nu\lambda, \\ z, \quad when\ |z| \geq \nu\lambda. \end{cases}$$

When $\nu < \kappa$, the solution to (4.12) is given by $\gamma^* = zI(|z| \geq \nu\lambda)$.

See Appendix C for the proof. The expressions for the cases $\nu < \kappa$ and $\nu - 1 < \kappa \leq \nu$ can be combined; we did not do this to simplify the proof. When $\kappa < \nu - 1$, the optimal solution of (4.12) is a continuous function of z, but this is not true for $\kappa > \nu - 1$. Thus, SCAD may not be the most appropriate choice for the order selection of mixture models. At the same time, because $\kappa = ua(\phi)$ is usually very small as in (4.11), the adverse effect is likely minor in applications. We focus on the algorithm and leave this issue for future research.

Proposition 4.1 leads to a thresholding operation for (4.12). Taking SCAD for example, shrinkage occurs whenever the generic value of $|z|$ is lower than some threshold level, say $\kappa\lambda$. This feature leads to the sparse $\boldsymbol{\eta}$. Suppose the initial vector $\boldsymbol{\eta}^{(0)}$ contains K components. Each incident of $\hat{\eta}_j = 0$ for some j reduces the order of the candidate mixing distribution by one. Because of this feature, the order of the MPLE is a direct output of the algorithm. No ad hoc steps are required.

4.4 Tuning strategies

In this section, we discuss some ITD implementation issues.

4.4.1 Choice of u

The tuning parameter u does not have a statistical implication. It controls by how much the iteration is directed by the gradient of $Q(\boldsymbol{\eta})$. We may therefore try to use a large u value. However, the inequality in the proof of Theorem 1 indicates that the objective function is guaranteed to decrease after each iteration only when u is small enough. Because $\kappa = ua(\phi)$ and the size of this largely controls the thresholding level, a large u also helps to speed up the algorithm.

For normal mixtures, we have $b''(\theta) = 1$ so we may choose $u = (n\tau_1)^{-1}$, the largest value allowed to ensure the monotonicity of $Q(\boldsymbol{\eta}^{(m)})$. For Poisson mixtures, we may choose u adaptively according to the value of $\boldsymbol{\eta}^{(m)}$ to guarantee monotonicity. Once the sequence $\boldsymbol{\eta}^{(m)}$ settles down, a stable value of u is used.

To accelerate the algorithm, we employ the following tuning strategy. Let $w^* = \max_k \sum_{i=1}^n w_{ik}$ and Γ_0 be the submatrix of Γ with columns corresponding to the nonzero entries of $\boldsymbol{\eta}^{(0)}$. Denote by τ_1^* the maximum eigenvalue of $\Gamma_0^T\Gamma_0$ and let $\tau_2^* = \max_k b''(\theta_k^{(0)})$. At each thresholding iteration, we first initiate an tentative large value for u, i.e., $u^{(m)} = u^* = 6/(w^*\tau_1^*\tau_2^*)$, and then check whether or not $Q(\boldsymbol{\eta}^{(m+1)}) < Q(\boldsymbol{\eta}^{(m)})$. When the inequality is violated, we reduce u to $u^{(m)} = 0.5u^*$, and we repeat this reduction procedure until $Q(\boldsymbol{\eta}^{(m+1)}) < Q(\boldsymbol{\eta}^{(m)})$ is satisfied. It is clear that the adaptive procedure does not alter the convergence of the algorithm.

4.4.2 Choice of λ

The tuning parameter λ in $p_\lambda(\cdot)$ controls the shrinkage in $\boldsymbol{\eta}$, and therefore it is crucial for the resulting order of the finite mixture model. When $\lambda = 0$, the MPLE of (4.4) places no penalty and thus results in the largest order. As λ increases from zero to infinity, the MPLE gradually reduces the order for the mixture model to one. The choice of λ relates to the trade-off between parsimony and goodness-of-fit.

Chen and Khalili [2008] prove that when λ has a certain asymptotic order, the SCAD regularization method is consistent when the sample size increases to infinity. However, the asymptotic result does not provide a specific recommendation for a λ-value in practice. Instead, a cross-validation (CV) procedure is recommended.

Specifically, let the full data set $Y = \{y_1, \ldots, y_n\}^T$ be divided into R nonoverlapping subsets, say Y_r with size n_r for $r = 1, \ldots, R$ and $\sum_{r=1}^{R} n_r = n$. Let $Y - Y_r$ be the subset with Y_r removed from Y. Let $\hat{\boldsymbol{\Psi}}_{\lambda,-r}$ be the MPLE of the model parameter $\boldsymbol{\Psi}$ based on $Y - Y_r$. The CV procedure selects a λ that minimizes

$$\text{CV}(\lambda) = -\sum_{r=1}^{R} l_{n_r}(\hat{\boldsymbol{\Psi}}_{\lambda,-r}; Y_r),$$

where $l_{n_r}(\boldsymbol{\Psi}; Y_r)$ is the log-likelihood function based on Y_r. The choice of λ with $r = 1$ worked well in Chen and Khalili's simulation studies. At the same time, the CV combined with an exhaustive search of λ makes the whole procedure computationally intensive. A more efficient tuning strategy is desirable in practice. We implemented the algorithm of Chen and Khalili [2008] with a 20-fold CV (Zhang [1993]) to reduce the computational burden.

In the literature, BIC-type criteria are investigated for choosing the tuning parameter in regularization methods for variable selection (Wang et al. [2007]). In such methods, the candidate choices of λ are evaluated by the likelihood scores of the MPLE with penalties proportional to the model complexity. The λ yielding the best score is then selected. Compared with the CV, this method is computationally cheaper and often leads to a satisfactory selection result. Therefore, we define a revised version of BIC as a function of λ by

$$\text{RBIC}(\lambda) = -2l_n(Y; \hat{\Psi}_\lambda) + \gamma\tau(\hat{\Psi}_\lambda)\log n,$$

where $\hat{\Psi}_\lambda$ is the MPLE with regularization parameter value λ and $\tau(\hat{\Psi}_\lambda)$ is the number of parameters in this mixing distribution. With $\gamma = 1$, we obtain the traditional BIC. In the simulation, we used $\gamma = 0.25$ and 0.5. We chose a smaller than usual value partly for its superior performance in our simulation studies and partly because an extra component in mixture models usually leads to a single extra degree of freedom in the limiting distribution of likelihood-based test statistics (Mengersen et al. [2011]). Once the value of γ is chosen, the RBIC selects a λ value that minimizes RBIC(λ) and thus suggests a corresponding order for the finite mixture model.

In applications, several trial runs are helpful to identify the proper range for λ. We adopted a data-driven searching strategy in our numerical studies. We begin with a large value $\lambda = \lambda^*$ that leads to a homogeneous model with order 1. The search proceeds by bisectionally reducing λ^* toward zero, such that the orders of the resulting models increase gradually. Grid searches are used between these λ values to obtain models with intermediate orders. The efficiency of this strategy has been observed in our simulation studies.

4.4.3 Choice of C_K

The constant C_K in the first penalty of (4.4) controls the penalization for small mixing proportions. Its value influences the precision of the corresponding MPLE of the model parameters. However, the effect of C_K has repeatedly been found to be minor for the order selection. Chen et al. [2001] and Chen and Khalili [2008] suggested that if the component parameters θ_k are restricted within $[-M, M]$ or $[M^{-1}, M]$ for some large M, then an appropriate choice of C_K is $\log M$. In our numerical studies, we set $C_K = \log\log y^*$ for Poisson mixture models with $y^* = \max_i\{y_i, i = 1, \dots, n\}$. For normal mixture models, we set C_K to the logarithm of the maximum absolute observation based on the standardized data (see Section 4.5).

4.5 Numerical studies

We assess the performance of ITD by Monte Carlo simulation examples. In particular, we implement MSCAD with both LQA and ITD to examine their computa-

Table 4.1: Mixing proportions and component parameters in simulation models

Normal mixture models

Model	(π_1, μ_1)	(π_2, μ_2)	(π_3, μ_3)	(π_4, μ_4)	(π_5, μ_5)	(π_6, μ_6)
1	(0.5, 0)	(0.5, 3)				
2	(0.3, 0)	(0.7, 3)				
3	(0.2, 0)	(0.4, 4)	(0.4, 6)			
4	(1/4, 0)	(1/4, 3)	(1/4, 6)	(1/4, 9)		
5	(1/4, 0)	(1/4, 2)	(1/4, 5)	(1/4, 8)		
6	(0.3, 0)	(0.2, 2)	(0.3, 4)	(0.2, 6)		
7	(1/6, 0)	(1/6, 3)	(1/6, 6)	(1/6, 9)	(1/6, 12)	(1/6, 15)
8	(1/6, 0)	(1/6, 2)	(1/6, 4)	(1/6, 6)	(1/6, 9)	(1/6, 12)
9	(1/6, 0)	(1/6, 2)	(1/6, 4)	(1/6, 7)	(1/6, 9)	(1/6, 11)
10	(1/6, 0)	(1/6, 2)	(1/6, 4)	(1/6, 6)	(1/6, 8)	(1/6, 10)

Poisson mixture models

Model	(π_1, μ_1)	(π_2, μ_2)	(π_3, μ_3)	(π_4, μ_4)
1	(1/2, 1)	(1/2, 3)		
2	(1/5, 1)	(4/5, 3)		
3	(4/5, 1)	(1/5, 3)		
4	(1/4, 1)	(1/4, 4)	(1/4, 12)	(1/4, 20)
5	(0.1, 1)	(0.2, 4)	(0.3, 12)	(0.4, 20)
6	(0.4, 1)	(0.3, 4)	(0.2, 12)	(0.1, 20)

tional efficiency. The performances of a number of tuning strategies for λ are also compared. The algorithm has been implemented in the R software.

4.5.1 Simulations

We generated data from the normal and Poisson mixture models. The component mean $\boldsymbol{\mu}$ and mixing proportions $\boldsymbol{\pi}$ are given in Table 4.1. For the normal mixture models, we set the common component variance $\sigma^2 = 1$. To make the selection invariant on the data scale, we standardized the observations from normal mixtures such that the corresponding sample mean was zero and the sample standard deviation was three.

For each simulated data set, MSCAD was used for the order selection with a common upper bound $K = 12$. The EM procedure used by both LQA and ITD was implemented with the same initial settings as in Chen and Khalili [2008]. That

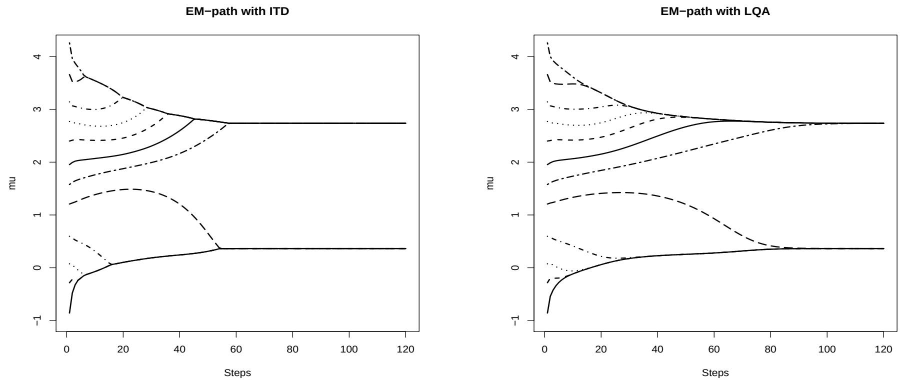

Figure 4.1: EM-paths of MSCAD for normal mixture model 1 with $n = 100$.

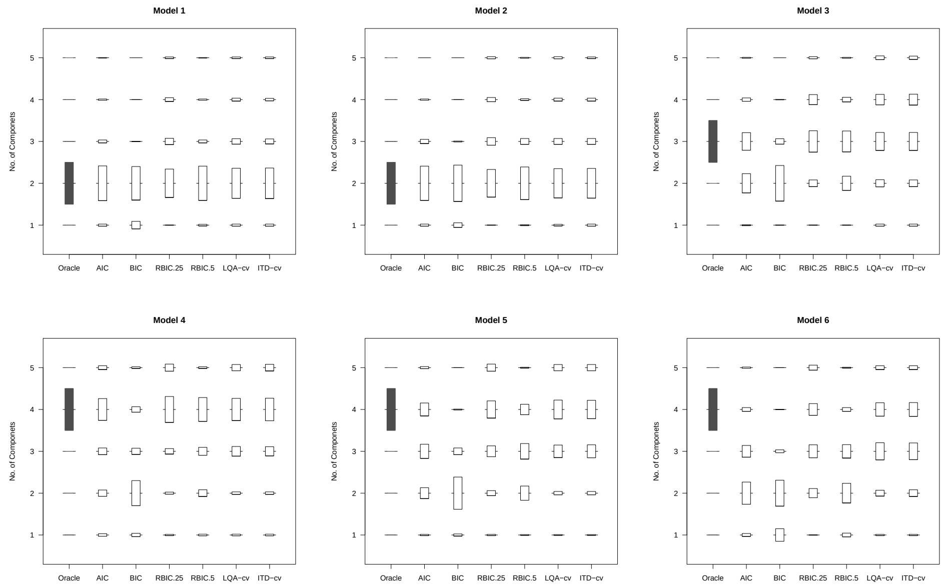

Figure 4.2: Selection results for normal mixture models (1–6), n=100.

is, the initial values of the component means were chosen as the $100(k-1/2)K\%$ sample quantiles and $\pi_k^{(0)} = 1/K$ for $k = 1, \ldots, K$. For the normal mixtures, we set the initial value of the component variance to the sample variance based on the observations trimmed at the 25% and 75% sample quantiles. For LQA, we merged two subpopulations if their component means were within 10^{-3} at the convergence

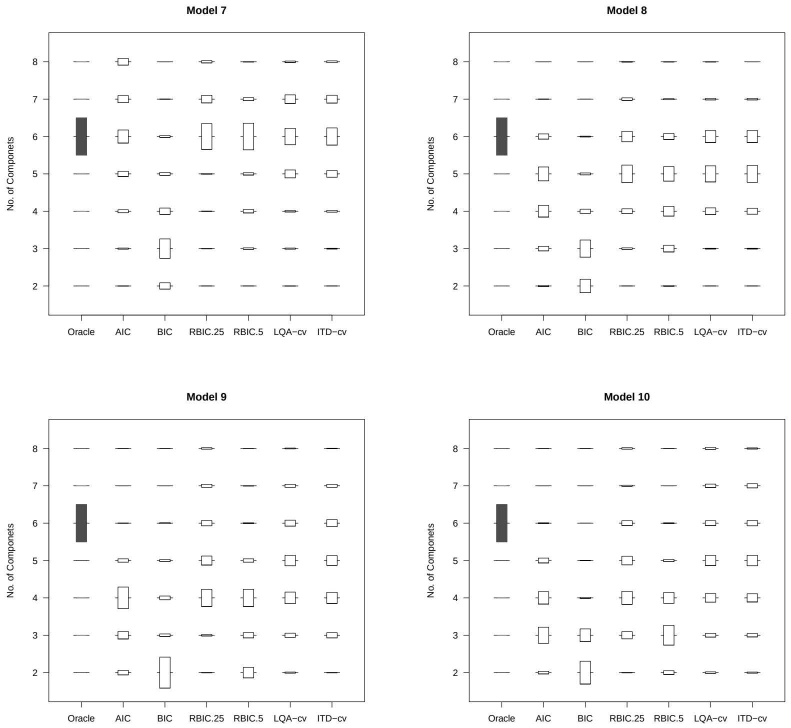

Figure 4.3: Selection results for normal mixture models (7–10), n=300.

of the algorithm.

As a concrete illustration, we plot in Fig. 4.1 the EM-iteration paths of $\boldsymbol{\theta}^{(t)}$ based on ITD and LQA for one data set generated from a normal mixture. Compared with EM-LQA, the paths of the EM-ITD procedure are sharper and require significantly fewer iterations to converge. We find that ITD saves about 40% of the computational time compared to LQA over our simulation studies. Although there is no guarantee, ITD and LQA usually converge to similar limiting points in our examples.

We summarize the simulation results in terms of how often various orders are selected based on 500 data sets generated from each model with sample sizes $n = 100, 300$. We use scaled bar plots to denote the frequency of the orders selected by MSCAD against methods and tuning strategies. The length of each

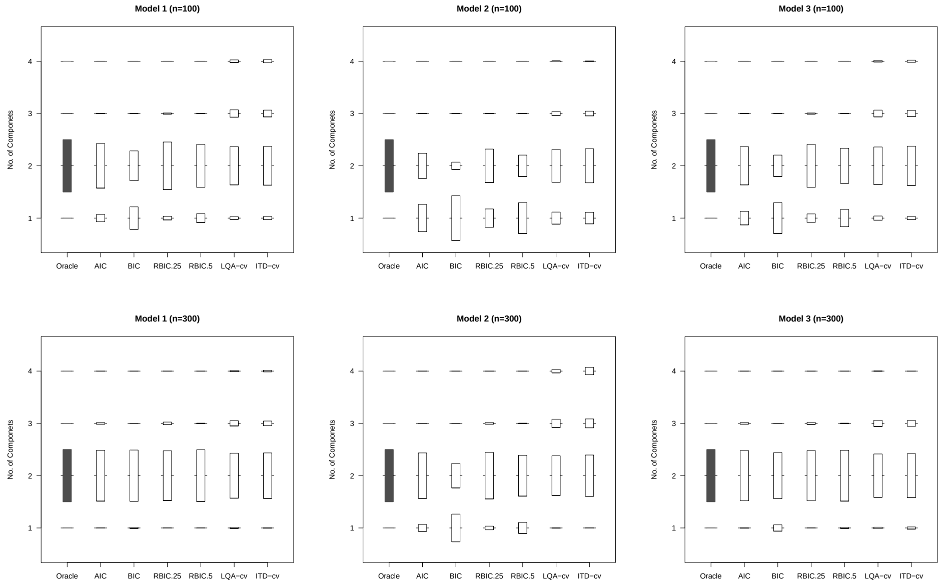

Figure 4.4: Selection results for Poisson mixture models (1–3), n=(100, 300).

bar is proportional to the selection frequency. As a benchmark, the true orders are marked by a black bar in each plot. We labeled the x-axes as follows: RBIC$_{.25}$ and RBIC$_{.5}$ are obtained by ITD with scaled RBIC; LQA-cv and ITD-cv are obtained by 20-fold CV based on LQA/ITD. We have also included the results of the popular AIC and BIC methods as suggested by Leroux [1992]. The simulation results are reported in Figs. 4.2–4.5.

Although it is not the goal of our simulation, these figures reaffirm the effectiveness of MSCAD for order selection. The performance of MSCAD is not affected by its implementation. This is reflected by the nearly identical columns for LQA-cv and ITD-cv in all the plots.

In terms of the selection frequency for the true mixture order, AIC and BIC do not perform well in most cases. In general, BIC grossly underestimates the true order of the model because of its stringent penalty on the model complexity. AIC is relatively liberal, but its performance is not satisfactory for models with complex structure.

The remaining results all relate to MSCAD but are tuned or implemented dif-

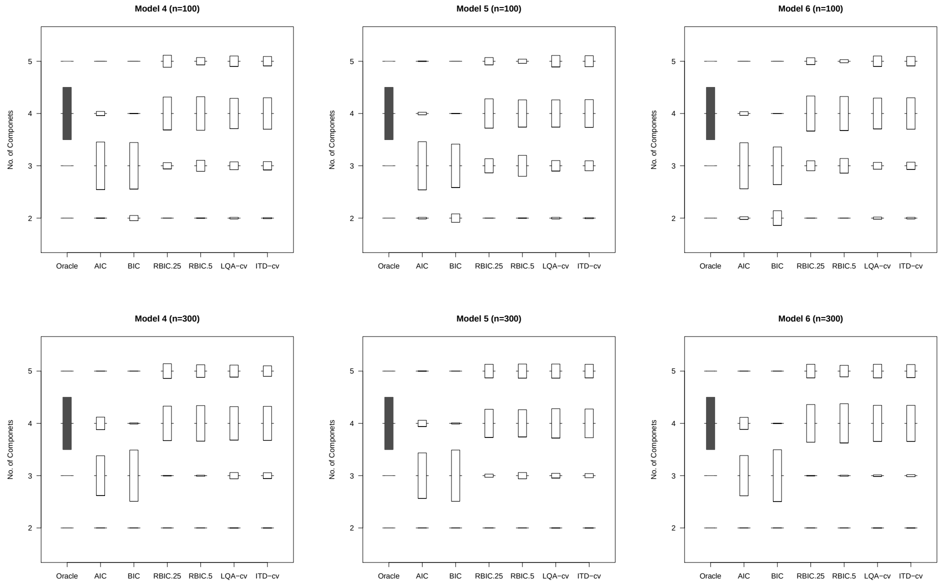

Figure 4.5: Selection results for Poisson mixture models (4–6), n=(100, 300).

ferently (i.e., RBIC$_{.25}$, RBIC$_{.5}$, ITD-cv, LQA-cv). Based on these plots, we see that RBIC$_{.5}$ tends to underestimate the true order. The other methods are comparable. If we factor in the computational savings, RBIC$_{.25}$ has the best performance. Also, the ITD-based methods perform better than the LQA because of the numerical advantages pointed out earlier.

4.5.2 Examples

We now illustrate the use of the ITD algorithm for order selection in real applications. We chose to implement MSCAD as in the simulation studies. Our first example is from a well-known astronomical data set, which consists of the velocities of 82 galaxies moving away from our own galaxy (available in Table 1 of Postman et al. [1986]; see the histogram in Fig. 4.6). The multimodality of the velocities may indicate the presence of super-clusters of galaxies surrounded by large voids, with each mode representing a cluster as it moves away at its own speed (see Roeder [1990] for more background). We may hence model the values of observed velocities as a random sample from a finite mixture of normal distributions with a

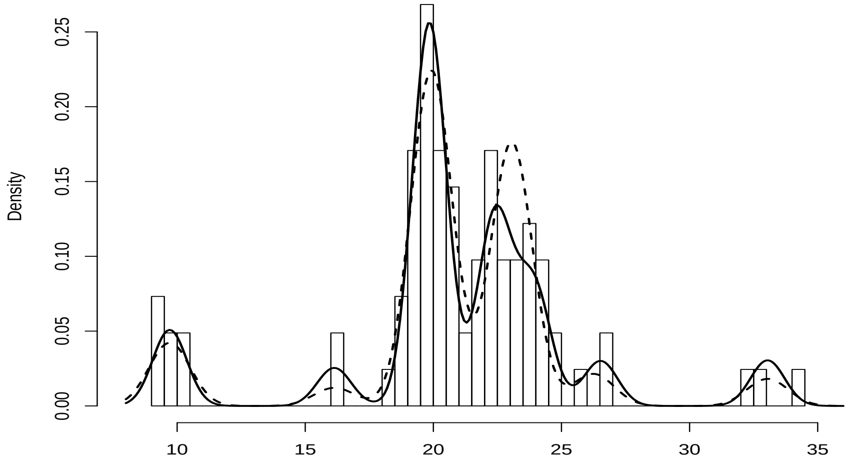

Figure 4.6: Histogram of galaxy data. Solid curve: density of seven component models selected by MSCAD. Dashed curve: density of six component models selected by AIC/BIC.

common variance. Using a Bayesian approach, Richardson and Green [1997] conclude that the number of distributions ranges from five to seven, which provides support for the existence of super-clusters.

MSCAD-ITD is used to reanalyze the galaxy data with $K = 12$. RBIC with $\gamma = 0.25$ was used to tune the parameter λ. The outcome is a seven-component model; the parameter estimates are given in Table 4.2 and Fig. 4.6 shows the fitted density function.

As a side note, AIC/BIC leads to a model with six components; see Table 4.2 and Fig. 4.6. It also provides a good description of the multimodal structure of the galaxy data.

Our next example is from Leroux and Puterman [1992], who discussed a study of breathing and body movements in fetal lambs. The number of movements of a fetal lamb was recorded for 240 consecutive 5-second intervals. These counts are overdispersed with a ratio of variance to mean estimated at 1.83. It is well known

Table 4.2: Parameter estimates for galaxy data. $\hat{\mu}_k$, $\hat{\pi}_k$, $\hat{\sigma}$ denote estimated component mean, mixing proportion, and common component standard deviation ($k = 1, 2, \dots, 7$).

Method	$(\hat{\mu}_1, \hat{\pi}_1)$	$(\hat{\mu}_2, \hat{\pi}_2)$	$(\hat{\mu}_3, \hat{\pi}_3)$	$(\hat{\mu}_4, \hat{\pi}_4)$	$(\hat{\mu}_5, \hat{\pi}_5)$	$(\hat{\mu}_6, \hat{\pi}_6)$	$(\hat{\mu}_7, \hat{\pi}_7)$	$\hat{\sigma}$
MSCAD	(9.7, .08)	(16.1, .04)	(19.8, .43)	(22.4, .21)	(23.9, .14)	(26.5, ,05)	(33.1, .05)	.66
AIC/BIC	(9.7, .09)	(16.2, .02)	(19.9, .45)	(23.1, .36)	(26.3, .04)	(33.0, .04)	(- -.-, .- -)	.81

that overdispersion can often be explained by population heterogeneity. Hence, a Poisson mixture model was used with the number of components selected by AIC/BIC. Using MSCAD implemented via ITD and an initial value of $K = 12$, we select a two-component model. Not surprisingly, this order is the same as that of the previous analysis. We summarize the resulting fits in Table 4.3. Because of its built-in measure to prevent type-I overfitting, MSCAD gives more balanced mixing proportions. Clustering more observations into the second component also leads to lower component means.

Table 4.3: Parameter estimates for lamb data. $\hat{\mu}_k$, $\hat{\pi}_k$ denote estimated component mean and mixing proportion ($k = 1, 2$).

Method	$(\hat{\mu}_1, \hat{\pi}_1)$	$(\hat{\mu}_2, \hat{\pi}_2)$
MSCAD	(.227, .920)	(1.88, .080)
AIC/BIC	(.230, .939)	(2.32, .061)

4.6 Summary and conclusions

In this chapter, we have developed an iterative thresholding descent algorithm (ITD) for PLM-based order selection methods in finite mixture models. The new algorithm avoids directly solving the original multivariate optimization problem and efficiently leads to the sparse MPLE of the component parameters. We established the algorithmic convergence of ITD under mild conditions. The efficiency of ITD is well supported by our numerical studies.

In applications, we need to specify an upper bound K for the MSCAD-based analysis. One should always use a sufficiently large K to avoid underestimation. However, too large a K often leads to slow convergence. In the examples we examined, the outcomes are not sensitive to the choice of K within a plausible range. We do not recommend setting K below 5 or above 15 in most applications. One needs strong empirical evidence or a large sample size (say above 300) to go outside this range.

Chapter 5

Summary and Future Work

5.1 Summary of the dissertation

In this dissertation, we have developed approaches based on PLMs to address the issues of feature selection arising in several application fields.

In Chapter 2, we addressed feature selection for ultra-high-dimensional data, where the number of features (covariates) is larger than the sample size. To facilitate the selection process, we proposed a novel screening approach to reduce the dimensionality of the data before the application of the PLM. The new method is motivated by the idea of the sparsity-restricted maximum likelihood estimator and can be efficiently implemented through a thresholding-based algorithm. Compared with the existing approaches which screen features based on the marginal correlations between the covariates and response, our method accounts for more joint effects between the covariates and thus can be more reliable in applications. We further established the consistency of our method in an ultra-high-dimensional setup and demonstrated its excellent performance through numerical examples.

In Chapter 3, we addressed variable selection for complex survey data, where the observations are intrinsically dependent because of the without-replacement sampling plan. To avoid a distorted conclusion caused by the biased samples, we proposed a penalized pseudo-likelihood approach to incorporate the survey weights into the selection process. A pseudo-likelihood-based BIC was further suggested to select the corresponding tuning parameter. We demonstrated the asymptotic consistency of our selection method in a joint randomization framework. The decent performance and potential benefits of our method were demonstrated in simulation studies.

In Chapter 4, we addressed order selection in finite mixture models for heterogeneous data. A thresholding-based algorithm was proposed for the implementation of PLMs in such applications. The new algorithm transforms the original multivariate objective function into a sum of univariate functions and efficiently leads to a sparse solution without ad hoc steps. We established the convergence of the new algorithm and illustrated its efficiency through both simulations and real-data examples.

5.2 Future directions

We have demonstrated that the PLM is an attractive technique for statistical learning with a wide range of applications. In the remainder of this chapter, we present several possible directions for future research.

5.2.1 Variable selection in model-based clustering

Clustering is a fundamental data-analysis tool which assigns similar objects to groups. When the data dimension is high, there are many noise variables that may mask the underlying clustering structures. Specifically, given the p-dimensional observations $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$ for $i = 1, \dots, n$, we aim to group the observations into a few clusters such that observations in the same cluster are more similar to each other than to those in different clusters. When p is large, it is likely that many entries in $\boldsymbol{x}_i$ are not relevant to the clustering: including them in the analysis introduces noise which might hide the heterogeneous structure of the data. Removing these “noise” variables is essential for the clustering of high-dimensional data.

Of the many clustering methods, model-based clustering (McLachlan and Peel [2002], Zhong and Ghosh [2003]) is appropriate for variable selection methods. These approaches assume that the data are generated from a finite mixture distribution with each component corresponding to a cluster. In this framework, the removal of “noise” variables can be performed by a model selection procedure, in which the shrinkage idea of PLM is helpful.

Specifically, suppose that, prior to the clustering, the data are standardized such that each entry $X_j = (x_{1j}, \ldots, x_{nj})^T$ has sample mean 0 and sample variance 1. In some applications, it might be appropriate to assume that the data are independently generated according to a p-variate normal mixture density

$$f(\boldsymbol{x}; \boldsymbol{\Omega}) = \sum_{k=1}^{K} \pi_k h(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{5.1}$$

where $\pi_k > 0$ is the mixing proportion such that $\sum_{k=1}^{K} \pi_k = 1$, $h(.)$ denotes the p-variate normal density with mean vector $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kp})$ and covariance matrix $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\Omega} = \{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), k = 1, \ldots, K\}$ denotes the unknown parameters. Clearly, if some components of $\boldsymbol{\mu}_k$ are small, the corresponding entries of the data are not informative for the clustering, at least in terms of location. Employing the shrinkage idea, we may use the PLM to automatically set small components of $\boldsymbol{\mu}_k$ to zero and thus remove the irrelevant entries from the data. We could address this by maximizing

$$Q(\boldsymbol{\Omega}) = l(\boldsymbol{\Omega}) - \varphi(\boldsymbol{\Omega}) - \sum_{j=1}^{p} \phi_\lambda(\|\boldsymbol{\mu}_{(j)}\|_2), \tag{5.2}$$

where $l(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \log f(\boldsymbol{x}_i; \boldsymbol{\Omega})$ is the log-likelihood, $\varphi(.)$ and $\phi_\lambda(.)$ are two penalty functions, and

$$\|\boldsymbol{\mu}_{(j)}\|_2 = \sqrt{\sum_{k=1}^{K} \boldsymbol{\mu}_{kj}^2}$$

is the Euclidean norm on $\boldsymbol{\mu}_{(j)} = (\mu_{1j}, \ldots, \mu_{kj})$ for $j = 1, \ldots, p$. In the normal mixture cases, the first penalty $\varphi(.)$ is generally required to guarantee the existence of the maximizer of (5.2) (Chen et al. [2008]), while the second penalty $\phi_\lambda(.)$ can lead to a sparse estimate of the mean parameter. Placing $\|\boldsymbol{\mu}_{(j)}\|_2$ in $\phi_\lambda(.)$ helps us to set $\boldsymbol{\mu}_{(j)} = 0$, so that the corresponding entry can be regarded as irrelevant to the clustering. A similar idea has been adopted by Yuan and Lin [2006] for a group LASSO problem. The optimization problem of (5.2) can be similarly cast into an EM-based procedure as discussed in Chapter 4.

As in the standard model-based clustering, one could further fit (5.1) by maximizing (5.2) with a series of K values and use a selection criterion (e.g., BIC) to determine the optimal number of clusters. Both the theoretical and computational aspects of this procedure require further research.

5.2.2 Support vector machine with nonconvex penalty

Classification is a basic task in pattern recognition and data mining. Kernel-based classification technologies have attracted much attention, because of their elegant interpretation and highly competitive performance. The support vector machine (SVM) is among the most popular classification approaches in this category (Cristianini and Shawe-Taylor [2000]); it learns the classification rule from the informative margins of the training data.

In a binary classification problem, the response y is either +1 or -1 (the classification labels), and a classification rule T is a mapping from the feature vector $\boldsymbol{x}$ to $\{+1, -1\}$. Given the training set $(y_i, \boldsymbol{x}_i)$ for $i = 1, \ldots, n$, we must find a discriminant function $f(.)$ so that any new input $\boldsymbol{x}$ can be correctly classified with $f(\boldsymbol{x})$ based on T. The kernel-based methods generally use the rule $T = \text{sign}[f(\boldsymbol{x})]$ and the following $f(.)$:

$$f(\boldsymbol{x}; \boldsymbol{w}) = \sum_{i=1}^{n} w_i k(\boldsymbol{x}, \boldsymbol{x}_i) + w_0, \tag{5.3}$$

where $k(.)$ is a user-specified kernel function such as the polynomial or Gaussian kernel, and $\boldsymbol{w} = (w_0, \ldots, w_n)$ is the weight vector to be estimated. Given a loss function $L(.)$, the kernel method (5.3) leads to the following optimization scheme

$$\min_{\boldsymbol{w}} \left\{ n^{-1} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i; \boldsymbol{w})) + \phi_\lambda(\boldsymbol{w}) \right\}, \tag{5.4}$$

where $\phi_\lambda(.)$ is a penalty function index with tuning parameter λ.

With the L_2 penalty $\phi_\lambda(\boldsymbol{w}) = \lambda \|\boldsymbol{w}\|_2^2$ and the hinge loss, i.e.,

$$L(y, f) = (1 - y \cdot f)_+ = \max\{0, 1 - y \cdot f\},$$

(5.4) leads to the typical L_2-SVM, which has been applied to a number of classification problems. Solid statistical learning theory for the use of L_2-SVM has been developed (Lin [2002]). As in ridge regression, the L_2 penalty helps to control the model complexity to prevent overfitting.

With the development of sparse regularization methods (e.g., LASSO), the L_1 penalty has been used in SVM (Zhu et al. [2003]); this leads to a sparse estimate of $\boldsymbol{w}$ in the discriminant function (5.3). The so-called L_1-SVM can result in a more compressive rule and thus provide a clearer interpretation. Friedman et al. [2004] showed that the L_1-SVM performs better if the underlying model is sparse, while the L_2-SVM performs better if most of the observations contribute to the response. Because of the promising theoretical properties of nonconvex penalties, it is natural to explore whether using them in (5.4) would further benefit the SVM. Such a nonconvex-penalized SVM may help to reduce the biased estimates of L_1, while maintaining the desirable feature of sparsity in the classification rule. It would be interesting to continue this investigation.

5.2.3 Some other issues

Variable selection via grouping

In regression analysis, one difficulty for variable selection is the collinearity between the covariates. In ultra-high-dimensional situations, even when all the covariates are ideally independent, the maximum sample correlation between covariates can still be high (Fan and Lv [2008]), which makes it hard to detect the truly influential covariates. Therefore, one might consider grouping all the candidate covariates according to their sample correlations and then treating each group as a independent predictor. In other words, covariates in the same group are considered to represent some factor that may be associated with the response. We could then use the penalized likelihood idea to choose the relevant groups.

One-step thresholding-based procedure

The thresholding-based algorithm provides a simple and efficient way to solve the numerical problem of nonconvex PLMs. However, because of the nonconcavity of the objective function, multiple initial values are often needed in its implementation. It would be beneficial to develop a data-driven strategy for initializing the thresholding-based procedure. A good initial value often leads to fewer iteration steps and guarantees the statistical properties of the estimate. In Chapter 2, we have tentatively revealed the efficiency of the LASSO-based initializing strategy. It would be beneficial if this method could be given a theoretical justification. Motivated by Zou and Li [2008], we are particularly interested in finding an initial setting such that a few iterations would suffice for variable selection/screening.

Tuning LAD-LASSO with high-dimensionality

LASSO is an computationally attractive approach to variable selection in high-dimension regression models. The $L1$-penalized least squares problem (1.9) has received much attention. However, the least squares method is known to be sensitive to outliers. An alternative approach is the least absolute deviation (LAD) method which can be robust to outliers. Wang et al. [2006] proposed the L_1-penalized LAD, i.e., LAD-LASSO, which is an efficient and robust selection approach. In particular, the idea of coordinate descent (CD) can be conveniently applied to the LAD-LASSO problem. LAD-LASSO therefore has great potential for high-dimensional noisy data. Despite its promising properties (Gao and Huang [2010]), the tuning parameter in LAD-LASSO needs to be specified in applications. The traditional AIC and BIC are designed for least square settings and may not be suitable for the LAD-LASSO problem. A feasible tuning method is therefore desired to realize the desirable properties of LAD-LASSO, especially for high-dimensional situations. EBIC (Chen and Chen [2008]) might provide an appropriate tuning strategy for LAD-LASSO and its variants.

Bibliography

H. Akaike. Information theory and an extension of the maximum likelihood principle. *In 2nd International Symposium on Information Theory*, 1:267–281, 1973. → pages 6, 70, 81

A. Antoniadis. Wavelet methods in statistics: Some recent developments and their applications. *Statistics Surveys*, 1:16–55, 2007. → pages 89, 90

A. Antoniadis and J. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96:939–967, 2001. → pages 90

J. Berkhof, I. Van Mechelen, and A. Gelman. A bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, 13:423–442, 2003. → pages 81

P. Bickel and D. Freedman. Asymptotic normality and the bootstrap in stratified sampling. *The Annals of Statistics*, 12:470–484, 1984. → pages 63

D. Binder. On the variances of asymptotically normal estimators from complex surveys. *International statistical Review*, 51:279–292, 1983. → pages 61

D. Binder and G. Roberts. *Chapter: Design-based and model-based methods for estimating model parameters*. Wiley Series in Survey Methodology, Chichester, 2003. → pages 59

T. Blumensath and M. Davies. Iterative thresholding for sparse approximations. *The Journal of Fourier Analysis and Applications*, 14:629–654, 2008. → pages 27

T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27:265–274, 2009. → pages 26, 32, 34, 36

D. Böhning. *Computer-Assisted Analysis of Mixtures and Applications: Meta Analysis, Disease Mapping, and Others*. Chapman & Hall/CRC, 2000. → pages 80

L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995. → pages 11, 20

T. Brijs, D. Karlis, G. Swinnen, K. Vanhoof, G. Wets, and P. Manchanda. A multivariate poisson mixture model for marketing applications. *Statistica Neerlandica*, 58:322–348, 2004. → pages 80

S. Canada. Survey on living with chronic diseases in canada 2009: User guide. *Supplementary documentation*, 2009. → pages 66

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52:489–509, 2006. → pages 26

I. Carrillo, J. Chen, and C. Wu. The pseudo-gee approach to the analysis of longitudinal surveys. *Canadian Journal of Statistics*, 38:540–554, 2010. → pages 62

H. Chen and J. Chen. The likelihood ratio test for homogeneity in finite mixture models. *Canadian Journal of Statistics*, 29:201–216, 2001. → pages 81

H. Chen, J. Chen, and J. Kalbfleisch. A modified likelihood ratio test for homogeneity in finite mixture models. *Journal of the Royal Statistical Society: Series B*, 63:19–29, 2001. → pages 93

J. Chen and Z. Chen. Extended bayesian information criterion for model selection with large model spaces. *Biometrika*, 95:759–771, 2008. → pages 41, 107

J. Chen and Z. Chen. Extended bic for small-n-large-p sparse glm. *Statistica Sinica*, 22:555–574, 2012. → pages 19, 33, 41, 42, 54, 55, 118

J. Chen and J. Kalbfleisch. Penalized minimum-distance estimates in finite mixture models. *Journal of the American Statistical Association*, 103: 1674–1683, 1996. → pages 81, 84

J. Chen and A. Khalili. Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103:1674–1683, 2008. → pages 24, 81, 82, 83, 85, 90, 92, 93, 94

J. Chen and J. Rao. Asymptotic normality under two-phase sampling designs. *Statistica Sinica*, 17:1047–1064, 2007. → pages 63

J. Chen, X. Tan, and R. Zhang. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18:443–465, 2008. → pages 104

P. Craven and G. Wahba. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979. → pages 18, 70

N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000. → pages 105

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004. → pages 82, 86

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39: 1–38, 1977. → pages 84

P. Dickson and J. Ginter. Market segmentation, product differentiation, and marketing strategy. *Journal of Marketing*, 51:1–11, 1987. → pages 1

D. Donoho. High-dimensional data analysis: the curses and blessings of dimensionality. *Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century*, 2000. → pages 25

D. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52: 1289–1306, 2006. → pages 26, 32

B. Efron. Correlation and large-scale simulations significance testing. *Journal of the American Statistical Association*, 102:93–103, 2007. → pages 30

B. Efron. Empirical bayes estimates for large-scale prediction problem. *Journal of the American Statistical Association*, 104:1015–1028, 2009. → pages 54

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004. → pages 20, 82, 86

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96: 1348–1360, 2001. → pages 2, 11, 12, 13, 17, 18, 21, 25, 38, 46, 64, 70, 81, 132

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*, 70: 849–911, 2008. → pages 22, 23, 26, 29, 30, 32, 106

J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010. → pages 25

J. Fan and J. Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57:5467–5484, 2011. → pages 17, 39

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32:781–813, 2004. → pages 17

J. Fan and R. Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38:3567–3604, 2009. → pages 30

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional variable selection: beyond the linear model. *Journal of Machine Learning Research*, 10: 1829–1853, 2009. → pages 30, 43, 44

I. Frank and J. Friedman. A statistical view of some chemonetrics regression tools. *Technometrics*, 35:109–148, 1993. → pages 12

J. Friedman, T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. Discussion of boosting papers. *The Annals of Statistics*, 32:102–107, 2004. → pages 106

J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1:302–332, 2007. → pages 20, 82, 86

W. J. Fu. Penalized regression: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998. → pages 11, 12, 20

W. Fuller. *Sampling Statistics*. Hoboken, New Jersey: Wiley, 2009. → pages 57, 74

X. Gao and J. Huang. Asymptotic analysis of high-dimensional lad regression with lasso. *Statistica Sinica*, 20:1495–1506, 2010. → pages 107

R. Gelber, J. Gaziano, J. Manson, J. Buring, and H. Sesso. A prospective study of body mass index and the risk of developing hypertension in men. *American Journal of Hypertension*, 20:370–377, 2007. → pages 78

J. Ghosh and P. Sen. On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In *In Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 789–806, 1985. → pages 81

V. Godambe and M. Thompson. Parameters of superpopulation and survey population: Their relationship and estimation. *International Statistical Review*, 54:127–138, 1986. → pages 61

J. Hájek. Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of Hungarian Academy of Science*, 5:361–375, 1960. → pages 62

J. Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35:1491–1523, 1964. → pages 63

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining*. Springer-Verlag, New York, 2 edition, 2009. → pages 25

H. Ishwaran, L. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96:1316–1332, 2001. → pages 81

G. Kalton. Models in the practice of survey sampling. *International Statistical Review*, 51:175–188, 1983. → pages 59

R. Kass and A. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995. → pages 65

A. Keller, P. Leidinger, A. Borries, A. Wendschlag, F. Wucherpfennig, M. Scheffler, H. Huwer, H. Lenhof, and E. Meese. mirnas in lung cancer - studying complex fingerprints in patient's blood cells by microarray experiments. *BMC Cancer*, 9:353–363, 2009. → pages 1

C. Keribin. Consistent estimation of the order of mixture models. *The Indian Journal of Statistics*, 96:1316–1332, 2000. → pages 81

K. Knight and W. J. Fu. Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28:1356–1378, 2000. → pages 15, 16

E. Korn and B. Graubard. *Analysis of Health Surveys*. Wiley, New York, 1999. → pages 56

P. S. Kott. A model-based look at linear regression with survey data. *The American Statistician*, 45:107–112, 1991. → pages 74

N. Kushmerich. Learning to remove internet advertisements. *In Proceeding of the 3rd International Conference on Autonomous Agents*, pages 175–181, 1999. → pages 1

B. Leroux. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20:1350–1360, 1992. → pages 81, 97

B. Leroux and M. Puterman. Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models. *Biometrics*, 48:545–558, 1992. → pages 99

M. Lesperance and J. Kalbfleisch. An algorithm for computing the nonparametric mle of a mixing distribution. *Journal of the American Statistical Association*, 87:120–126, 1992. → pages 83

P. Li and J. Chen. Testing the order of a finite mixture model. *Journal of the American Statistical Association*, 105:1084–1092, 2010. → pages 81

Y. Lin. Support vector machine and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002. → pages 106

B. G. Lindsay. The geometry of mixture likelihoods: A general theory. *The Annals of Statistics*, 11:86–94, 1983. → pages 83

S. L. Lohr and J. Liu. A comparison of weighted and unweighted analyses in the ncvs. *Journal of Quantitative Criminology*, 10:343–360, 1994. → pages 76

G. MacLachlan. On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture. *Applied Statistics*, 36:318–324, 1987. → pages 81

G. MacLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Hoboken, New Jersey: Wiley, 2 edition, 2008. → pages 85

S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993. → pages 34

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 2 edition, 1989. → pages 20, 26

J. McLachlan and D. Peel. *Finite mixture models*. Wiley, New York, 2002. → pages 103

N. Meindhausen and P. Buhlmann. Consistent neighborhood selection for high-dimensional graphs with the lasso. *The Annals of Statistics*, 34: 1436–1462, 2006. → pages 16

K. Mengersen, C. Robert, and D. Titterington. *Mixtures: Estimation and Applications*. Hoboken, New Jersey: Wiley, 1 edition, 2011. → pages 93

E. Molina and C. Skinner. Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics and Data Analysis*, 13: 395–405, 1992. → pages 61

J. F. Murray and K. Kreutz-Delgado. An improved focuss-based learning algorithm for solving sparse linear inverse problems. *Conference Record of the Thirty-Fifth Asilomar Conference on Signals, Systems and Computers*, pages 347–351, 2001. → pages 34

E. Ohlsson. Asymptotic normality for two-stage sampling from a finite population. *Probability Theory and Related Fields*, 81:341–352, 1989. → pages 63

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686, 2008. → pages 14

D. Pfeffermann. The role of sampling weights when modeling survey data. *International Statistical Review*, 61:317–337, 1993. → pages 57

D. Pfeffermann and D. J. Holmes. Robustness considerations in the choice of a method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society, Series A*, 148:268–278, 1985. → pages 74

M. Postman, J. Huchra, and M. Geller. Probes of large-scale structures in the corona borealis region. *The Astronomical Journal*, 92:1238–1247, 1986. → pages 98

M. Rahiala and T. Teräsvirta. Business survey data in forecasting the output of swedish and finnish metal and engineering industries: A kalman filter approach. *Journal of Forecasting*, 12:255–271, 1993. → pages 56

C. R. Rao and Y. H. Wu. A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76:369–74, 1989. → pages 9

S. Richardson and P. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B*, 59: 731–792, 1997. → pages 81, 99

K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85: 617–624, 1990. → pages 98

M. Royall. The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71:657–664, 1976. → pages 58

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464, 1978. → pages 6, 64, 81, 133

J. Shao. An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, 7:221–242, 1997. → pages 8

Y. She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, 3:384–415, 2009. → pages 22, 82, 86

Y. She. An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis*, In press, 2011. → pages 19, 70

R. Shibata. Asymptotic mean efficiency of a selection of regression variables. *The Annals of Statistics*, 35:415–423, 1983. → pages 8

M. Shoukri and G. MacLachlan. Parametric estimation in a genetic mixture model with application to nuclear family data. *Biometrics*, 50:128–139, 1994. → pages 80

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kanto, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002. → pages 54

M. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:111–147, 1974. → pages 18

J. Storey and R. Tibshirani. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003. → pages 30

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996. → pages 2, 11, 12, 18, 19, 25, 81

L. Tierney, R. Kass, and J. Kadane. Fully exponential laplace approximation to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84:710–716, 1989. → pages 65

J. Víšek. Asymptotic distribution of simple estimate for rejective, sampford and successive sampling. *Contributions to Statistics*, pages 263–275, 1979. → pages 63

H. Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104:1512–1524, 2009. → pages 19, 26, 45

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection via the lad-lasso. *Journal of Business and Economic Statistics*, 11: 1–6, 2006. → pages 107

H. Wang, R. Li, and C. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94:553–568, 2007. → pages 18, 41, 64, 92

H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society, Series B*, 71:671–683, 2009. → pages 19

W. Wolfson. *Analysis of labour force survey data for the information technology occupations 2000-2003*. WGW Services Ltd., Ottawa, Ontario, 2004. → pages 56

M. Woo and T. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 51:4379–4392, 2006. → pages 81

J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11:95–103, 1983. → pages 85

L. Yan, K. Liu, K. Matthews, M. Daviglus, T. Ferguson, and C. Kiefe. Psychosocial factors and risk of hypertension: The coronary artery risk development in young adults (cardia) study. *Journal of American Medical Association*, 290:2138–2148, 2003. → pages 78

Y. Yang. Can the strength of aic and bic be shared? a conflict between model identification and regression estimation. *Biometrika*, 92:937–950, 2005. → pages 8

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006. → pages 104

C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942, 2010. → pages 14, 25

P. Zhang. Model selection via multifold cross-validation. *The Annals of Statistics*, 21:229–231, 1993. → pages 92

Y. Zhang, R. Li, and C. Tsai. Regularization parameter selection via generalized information criterion. *Journal of the American Statistical Association*, 105: 312–323, 2010. → pages 64

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006. → pages 15, 16

S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of Machine Learning Research*, 4:1001–1037, 2003. → pages 103

J. Zhu, S. Rosset, T. Hastie, and R. Tibshirani. l_1-norm support vector machines. *Neural Information Processing Systems*, 16, 2003. → pages 106

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006. → pages 14, 16

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301–320, 2005. → pages 11, 14, 25

H. Zou and R. Li. One-step sparse estimation in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533, 2008. → pages 21, 22, 82, 86, 107

Appendix A

Supplementary Information for Chapter 2

We provide proofs of Theorems 2.1-2.4 in this appendix. Before getting into the detailed illustrations, we first state a technical lemma as follows.

Lemma A.1 *Let Y_i, $i = 1, \dots, n$ be independent random variables following exponential family distributions of form (2.1) with natural parameters $\theta_i \in \Theta$. Let μ_i and σ_i^2 denote the mean and variance of Y_i respectively. Let t_{ni}, $i = 1, \dots, n$ be real numbers such that*

$$\sum_{i=1}^{n} t_{ni}^2 \sigma_i^2 = 1, \quad \max_{1 \leq i \leq n} \{t_{ni}^2\} = O(n)$$

for some positive sequence $h_n = o(n)$. Then, for a sufficiently large n,

$$P\left(\sum_{i=1}^{n} t_{ni}(Y_i - \mu_i) > h_n\right) \leq \exp(-\frac{h_n^2}{3}).$$

Lemma A.1 states a useful property of exponential family for the illustration of theorems in this chapter, the proof of which can be found in Chen and Chen [2012].

A.1 Proof of Theorem 2.1

Theorem 2.1 shows that, under some regularity conditions on model (2.1) and the design matrix, the SMLE-based screening $\hat{s}$ retains all influential features in the true model s^* with probability tending to one. We illustrate the theorem by showing that asymptotically $\hat{s}$ falls into the collection of over-fitted models that contain s^* as a submodel. This is implied by the fact that the maximum likelihood score based on an over-fitted model is asymptotically greater than that of any under-fitted model with at least one feature in s^* excluded.

Proof: Let $\hat{\boldsymbol{\beta}}_s$ be the (unrestricted) MLE of $\boldsymbol{\beta}$ based on model s. The theorem is implied if $P\{\hat{s} \in \boldsymbol{S}_{+}^{k}\} \to 1$. Thus, it suffices to show that

$$P\{\max_{s \in \boldsymbol{S}_{-}^{k}} l_n(\hat{\boldsymbol{\beta}}_s) \geq \min_{s \in \boldsymbol{S}_{+}^{k}} l_n(\hat{\boldsymbol{\beta}}_s)\ \} \to 0, \tag{A.1}$$

as $n \to \infty$.

For any $s \in \boldsymbol{S}_{-}^{k}$, define $s' = s \cup s^* \in \boldsymbol{S}_{+}^{2k}$. Consider $\boldsymbol{\beta}_{s'}$ close to $\boldsymbol{\beta}_{s'}^{*}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^{*}\| = w_1 n^{-\tau_1}$ for some $w_1, \tau_1 > 0$. Clearly, when n is sufficiently large, $\boldsymbol{\beta}_{s'}$ falls into a small neighborhood of $\boldsymbol{\beta}_{s'}^{*}$, so that condition (T3) becomes applicable. Thus, by Taylor's theory, we have

$$\begin{aligned}
&l(\boldsymbol{\beta}_{s'}) - l(\boldsymbol{\beta}_{s'}^{*})\\
&= [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^{*}]^T S(\boldsymbol{\beta}_{s'}^{*}) - \frac{1}{2}[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^{*}]^T H(\tilde{\boldsymbol{\beta}}_{s'})[\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^{*}]\\
&\leq [\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^{*}]^T S(\boldsymbol{\beta}_{s'}^{*}) - \frac{c_1}{2} n \|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^{*}\|^2\\
&\leq w_1 n^{-\tau_1} \|S(\boldsymbol{\beta}_{s'}^{*})\| - \frac{c_1}{2} w_1^2 n^{1-2\tau_1},
\end{aligned} \tag{A.2}$$

where $\tilde{\boldsymbol{\beta}}_{s'}$ is an intermediate value between $\boldsymbol{\beta}_{s'}$ and $\boldsymbol{\beta}_{s'}^{*}$. Thus, for some generic positive constant c, we have

$$\begin{aligned}
&P\{l(\boldsymbol{\beta}_{s'}) - l(\boldsymbol{\beta}^*_{s'}) \geq 0\} \\
&\leq P\{\|S(\boldsymbol{\beta}^*_{s'})\| \geq cn^{1-\tau_1}\} \\
&\leq \sum_{j\in s'} P\{S_j^2(\boldsymbol{\beta}^*_{s'}) \geq ck^{-1}n^{2-2\tau_1}\} \\
&= \sum_{j\in s'} P\{S_j(\boldsymbol{\beta}^*_{s'}) \geq ck^{-\frac{1}{2}}n^{1-\tau_1}\} + \sum_{j\in s'} P\{-S_j(\boldsymbol{\beta}^*_{s'}) \geq ck^{-\frac{1}{2}}n^{1-\tau_1}\}.
\end{aligned} \tag{A.3}$$

Note that

$$S_j(\boldsymbol{\beta}^*_{s'}) = \sum_{i=1}^{n}[y_i - b'(\boldsymbol{x}^T_{is'}\boldsymbol{\beta}^*_{s'})]x_{ij} = \sum_{i=1}^{n}[y_i - \mu_i]x_{ij}.$$

Let $t_{ni} = x_{ij}(\sum_{i=1}^n x_{ij}^2\sigma_i^2)^{-1/2}$. By condition T4, we have $\sum_{i=1}^n t_{ni}^2\sigma_1^2 = 1$, $\max_i\{t_{ni}^2\} = O(n^{-1})$ and $n^{-1}\sum_{i=1}^n x_{ij}^2\sigma_i^2 \leq c_4$. Also, by condition T2, we have $k \leq w_2 n^{\tau_2}$. With these conditions, Lemma 1 gives the following probability inequality

$$\begin{aligned}
&P\{S_j(\boldsymbol{\beta}^*_{s'}) \geq ck^{-\frac{1}{2}}n^{1-\tau_1}\} \\
&\leq P\{\sum_{i=1}^{n} t_{ni}(y_i - \mu_i) > cn^{0.5(1-2\tau_1-\tau_2)}\} \\
&\leq c\exp(-n^{1-2\tau_1-\tau_2}).
\end{aligned} \tag{A.4}$$

By the same arguments, we also have

$$P\{-S_j(\boldsymbol{\beta}^*_{s'}) \geq ck^{-\frac{1}{2}}n^{1-\tau_1}\} \leq c\exp(-n^{1-2\tau_1-\tau_2}). \tag{A.5}$$

The inequalities (A.4) and (A.5) imply that, for some generic constant c,

$$P\{l(\boldsymbol{\beta}_{s'}) \geq l(\boldsymbol{\beta}^*_{s'})\} \leq ck\exp(-n^{1-2\tau_1-\tau_2}). \tag{A.6}$$

Consequently, by Bonferoni inequality and condition $1 - 2\tau_1 - 2\tau_2 > 0$, we have

$$
\begin{aligned}
& P\{\max_{s \in \boldsymbol{s}_-^k} l(\boldsymbol{\beta}_{s'}) \geq l(\boldsymbol{\beta}_{s'}^*)\} \\
& \leq \sum_{s \in \boldsymbol{s}_-^k} P\{l(\boldsymbol{\beta}_{s'}) \geq l(\boldsymbol{\beta}_{s'}^*)\} \\
& \leq ckp^k \exp(-n^{1-2\tau_1-\tau_2}) \leq c \exp(\tau_2 \log n + mn^{\tau_2} \log n - n^{1-2\tau_1-\tau_2}) = o(1).
\end{aligned}
$$

Because $l(\boldsymbol{\beta}_{s'})$ is concave in $\boldsymbol{\beta}_{s'}$, above result holds for any $\boldsymbol{\beta}_{s'}$ such that $\|\boldsymbol{\beta}_{s'} - \boldsymbol{\beta}_{s'}^*\| \geq w_1 n^{\tau_1}$.

For any $s \in \boldsymbol{S}_-^k$, let $\breve{\boldsymbol{\beta}}_{s'}$ be $\hat{\boldsymbol{\beta}}_s$ augmented with zeros corresponding to the elements in s'/s^*. By condition T2, it is seen that

$$
\|\breve{\boldsymbol{\beta}}_{s'} - \boldsymbol{\beta}_{s'}^*\| \geq \|\boldsymbol{\beta}_{s^*/s}^*\| \geq w_1 n^{\tau_1}.
$$

Consequently,

$$
P\{\max_{s \in \boldsymbol{S}_-^k} l(\hat{\boldsymbol{\beta}}_s) \geq \min_{s \in \boldsymbol{S}_+^k} l(\hat{\boldsymbol{\beta}}_s)\ \} \leq P\{\max_{s \in \boldsymbol{s}_-^k} l(\breve{\boldsymbol{\beta}}_{s'}) \geq l(\boldsymbol{\beta}_{s'}^*)\} = o(1)
$$

The theorem is proved. ■

A.2 Proof of Theorem 2.2

Theorem 2.2 shows that the IHT procedure (2.12) stepwise increases the likelihood score within the restricted model sparsity. Such an increment property further ensures the convergence of IHT to a local maximum of (2.7).

Proof: We first show the increment of $l_n(\boldsymbol{\beta}^{(t)})$. According to (2.9), we have

$$
\begin{aligned}
l(\boldsymbol{\beta}^{(t)}) &= h(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t)}) \\
&\leq h(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \\
&= l(\boldsymbol{\beta}^{(t+1)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|^2 \\
&\quad + \sum_{i=1}^{n} [b(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t+1)}) - b(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)}) - b'(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})(\boldsymbol{x}_i^T \boldsymbol{\beta}^{(t+1)} - \boldsymbol{x}_i^T \boldsymbol{\beta}^{(t)})].
\end{aligned}
$$

By the Taylor's expansion, for any θ and θ_0, we have

$$b(\theta) - b(\theta_0) - b'(\theta_0)(\theta - \theta_0) = \frac{1}{2}b''(\tilde{\theta})(\theta - \theta_0)^2$$

for some $\tilde{\theta}$ between θ and θ_0. Applying this expansion, we find

$$\begin{aligned} l(\boldsymbol{\beta}^{(t)}) &\leq l(\boldsymbol{\beta}^{(t+1)}) - \frac{u}{2}\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|^2 + \frac{1}{2}\rho^{(t)}\|\boldsymbol{X}\boldsymbol{\beta}^{(t+1)} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}\|^2 \\ &\leq l(\boldsymbol{\beta}^{(t+1)}) + \frac{1}{2}(\rho^{(t)}\rho_1 - u)\|\boldsymbol{\beta}^{(m+1)} - \boldsymbol{\beta}^{(m)}\|^2 \\ &\leq l(\boldsymbol{\beta}^{(t+1)}) \end{aligned}$$

as we have required $u > \rho^{(t)}\rho_1$. Apparently, the quality holds only if $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$. The increment property of AHT is hence proved.

Now let us move to the convergence of $\{\boldsymbol{\beta}^{(t)}\}$. By condition C3$'$, $l(\boldsymbol{\beta}_s)$ is a strict concave function over $\boldsymbol{\beta}_s$ for any s such that $\|s\|_0 \leq k$. Since $\|\boldsymbol{\beta}^{(t)}\|_0 \leq k$ and $l(\boldsymbol{\beta}^{(t)}) \leq l(\boldsymbol{\beta}^{(t+1)})$ for $t \geq 0$, the concavity of $l(.)$ implies that $\boldsymbol{\beta}^{(t)}$ stays in a bounded (compact) region. Thus, $\tau^* = \sup_t(\rho^{(t)})$ is finite based on the continuity of $b''(\cdot)$. By the similar arguments, we obtain

$$l(\boldsymbol{\beta}^{(t+1)}) - l(\boldsymbol{\beta}^{(t)}) \geq \frac{1}{2}(u - \rho^*\rho_1)\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\|^2.$$

which implies that $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\| \to 0$ due to the convergence of $\{l(\boldsymbol{\beta}^{(t)})\}$.

Also, the compactness implies that $\{\boldsymbol{\beta}^{(t)}\}$ has at least one limit point, say, $\tilde{\boldsymbol{\beta}} = \{\tilde{\beta}_1, \ldots, \tilde{\beta}_p\}^T$. Let $\{t_m\}$ be a subsequence such that $\lim_{m\to\infty}\boldsymbol{\beta}^{(t_m)} = \tilde{\boldsymbol{\beta}}$. Since $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\| \to 0$, we must also have $\lim_{m\to\infty}\boldsymbol{\beta}^{(t_m+1)} = \tilde{\boldsymbol{\beta}}$.

We next show that $\tilde{\boldsymbol{\eta}}$ is a local maximum of $l_n(\boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_0 \leq k$. By (2.9), we have

$$\boldsymbol{\beta}^{(t_m+1)} = \arg\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}; \boldsymbol{\beta}^{(t_m)}) \quad \text{subject to} \|\boldsymbol{\gamma}\|_0 \leq k.$$

¿From the bivariate continuity of $h_n(\boldsymbol{\xi}; \boldsymbol{\eta})$, letting $m \to \infty$, we get

$$\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\gamma}} h(\boldsymbol{\gamma}; \tilde{\boldsymbol{\beta}}) \quad \text{subject to} \|\boldsymbol{\gamma}\|_0 \leq k.$$

That is, $\tilde{\boldsymbol{\beta}}$ is a maximum of $h(\boldsymbol{\gamma}; \tilde{\boldsymbol{\beta}})$ with respect to $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\gamma}\|_0 \leq k$. We now split our discussion into following two cases.

Case 1: when $\|\tilde{\boldsymbol{\beta}}\|_0 < k$, let $\tilde{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\beta}} + u^{-1}\boldsymbol{X}^T[\boldsymbol{y} - b'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})]$. By (2.12), we have

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{H}(\tilde{\gamma}; k),$$

which implies that the kth largest (in absolute value) component of $\tilde{\boldsymbol{\gamma}}$ is zero. The definition of $\boldsymbol{H}$ then tells us

$$\tilde{\boldsymbol{\beta}} = [H(\tilde{\gamma}_1; 0), \ldots, H(\tilde{\gamma}_p; 0)]^T = \tilde{\boldsymbol{\gamma}},$$

which implies that $S(\tilde{\boldsymbol{\beta}}) = \boldsymbol{X}^T[\boldsymbol{y} - b'(\boldsymbol{X}\tilde{\boldsymbol{\beta}})] = 0$. Therefore, $\tilde{\boldsymbol{\beta}}$ is an unconstrained maximum of $l_n(.)$ which is also a local maximum subject to $\|\boldsymbol{\beta}\|_0 \leq k$.

Case 2: when $\|\tilde{\boldsymbol{\beta}}\|_0 = k$, we must have

$$\left.\frac{\partial h(\boldsymbol{\gamma}, \boldsymbol{\beta}')}{\partial \gamma_j}\right|_{\boldsymbol{\gamma}=\boldsymbol{\beta}'} = 0 \tag{A.7}$$

for any $j \in \{1, \ldots, p\}$ where $\tilde{\beta}_j \neq 0$. Let us rewrite $h(\boldsymbol{\gamma}; \tilde{\boldsymbol{\beta}})$ as

$$h(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}) = l(\boldsymbol{\gamma}) + T(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}) \tag{A.8}$$

so that

$$T(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}}) = -\frac{u}{2}\|\boldsymbol{\gamma} - \tilde{\boldsymbol{\beta}}\|_2^2 + \sum_{i=1}^{n}\{b(\boldsymbol{x}_i^T\gamma_j) - b(\boldsymbol{x}_i^T\tilde{\beta}_j) - b'(\tilde{\beta}_j)(\boldsymbol{x}_i^T\gamma_j - \boldsymbol{x}_i^T\tilde{\beta}_j)\}.$$

It is seen that

$$\left.\frac{\partial T(\boldsymbol{\gamma}, \tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\gamma}}\right|_{\boldsymbol{\gamma}=\tilde{\boldsymbol{\beta}}} = 0.$$

Hence, together with (A.7) and (A.8), this fact implies that $\partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta}_j$ is zero at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ for any $j \in \{1, \ldots, p\}$ where $\tilde{\beta}_j \neq 0$. Let $\tilde{\delta}$ be the minimum absolute value of non-zero components in $\tilde{\boldsymbol{\beta}}$. Then, for $\boldsymbol{\beta}$ such that $|\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j\| \leq 0.5\tilde{\delta}$, $\tilde{\boldsymbol{\beta}}$ must be a local maximum of $l(\boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_0 \leq k$.

With above arguments, we now justify convergence of $\{\boldsymbol{\beta}^{(t)}\}$ as follows. Note that, by condition T3$'$, there are finite number of local maximum of $l(\boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_0 \leq k$. Suppose $\boldsymbol{\beta}^{(t)}$ does not converge but has two limiting points, say $\tilde{\boldsymbol{\beta}}_1 \neq \tilde{\boldsymbol{\beta}}_2$. By what we have just proved, both are also local maxima of $l(\cdot)$. Let $\epsilon = \|\tilde{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_2\|$. Since $\|\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\| \to 0$, the distance between successive $\boldsymbol{\beta}^{(t)}$ goes to 0 as $t \to \infty$. Thus, there must be infinite many $\boldsymbol{\beta}^{(t)}$ which are at least $\epsilon/3$ distance from these two limiting points. The compactness condition hence implies that $\boldsymbol{\beta}^{(t)}$ has at least another limiting point which is also a local maximum of $l(\cdot)$. Let this one be called $\tilde{\boldsymbol{\beta}}_3$. Using the same argument, $\boldsymbol{\beta}^{(t)}$ must have additional limiting point $\tilde{\boldsymbol{\beta}}_4$ which implies $l(\cdot)$ has at least 4 stationary points. Repeatedly applying this logic implies $l(\cdot)$ has infinite many stationary points. This contradicts with the assumption, and therefore implies $\boldsymbol{\beta}^{(t)}$ converges. ■

A.3 Proof of Theorem 2.3

Theorem 2.3 shows that, for appropriate choices of penalty function, the SMLE-PLM procedure consistently estimates the model coefficients in the ultra-high dimensional GLM setup where $p \gg n$. Since SMLE-PLM is a two-step procedure, where the PLM is used on a screened model $\hat{s}$ from the ultra-high dimensional full model space, the randomness from both screening and PLM steps need to be accounted in the asymptotic analysis. We illustrate the theorem by showing that the MPLE converges (in probability) to the true model coefficients at a flat rate based on any $\hat{s}$ that might be obtained from the screening step.

Proof: Apparently, the screening consistency $P(s^* \subset \hat{s}) \to 1$ implies that $P(\hat{s} \in \boldsymbol{S}_+^k) \to 1$. To show the theorem, we investigate the performance of MPLE on all possible models over $\boldsymbol{S}_+^k$. Specifically, let $\boldsymbol{u} = (u_1, \ldots, u_k)^T$ be an arbitrary k-dimensional vector with $\|\boldsymbol{u}\|_2 = 1$ and $a_n = n^{-v}$. For any $s \in S_+^k$, let $\boldsymbol{\beta}_s = \boldsymbol{\beta}_s^* + a_n\boldsymbol{u}$, and thus,

$$\begin{aligned} &Q(\boldsymbol{\beta}_s) - Q(\boldsymbol{\beta}_s^*) \\ &= l(\boldsymbol{\beta}_s^* + a_n\boldsymbol{u}) - l_n(\boldsymbol{\beta}_s^*) - n\sum_{j\in s}[\phi_\lambda(\beta_j^* + a_n u_j) - \phi_\lambda(\beta_j^*)] \end{aligned} \quad \text{(A.9)}$$

When n is sufficiently large, $\boldsymbol{\beta}_s$ falls into a small neighborhood of $\boldsymbol{\beta}_s^*$, so that C3 becomes applicable. Following similar arguments in (A.2), we have the likelihood term in (A.9) bounded by

$$l(\boldsymbol{\beta}_s^* + a_n\boldsymbol{u}) - l(\boldsymbol{\beta}_s^*) \leq a_n\|S(\boldsymbol{\beta}_s^*)\| - \frac{c_1}{2}na_n^2. \tag{A.10}$$

At the same time, we have

$$\begin{aligned} n\sum_{j\in s}[\phi_\lambda(\beta_j^* + a_nu_j) - \phi_\lambda(\beta_j^*)] &\geq n\sum_{j\in s^*}[\phi_\lambda(\beta_j^* + a_nu_j) - \phi_\lambda(\beta_j^*)] \\ &= n\sum_{j\in s^*} a_nu_j\phi_\lambda^{'}(|\tilde{\beta}_j|)\text{sign}(\tilde{\beta}_j) \quad \text{(A.11)} \end{aligned}$$

for some $\tilde{\beta}_j$ between β_j^* and $\beta_j^*+a_nu_j$. By condition C2, we require $\min_{j\in s^*}|\beta_j^*| \geq w_1n^{-\tau_1}$. Thus, when n is large enough, we have $\tilde{\beta}_j > 0.5w_1n^{-\tau_1}$ for $j \in s^*$ and property P3 of $\phi_\lambda(.)$ implies that the penalty term in (A.9) is bounded by

$$n\sum_{j\in s^*} a_nu_j\phi_\lambda^{'}(|\tilde{\beta}_j|)\text{sign}(\tilde{\beta}_j) \geq -w_3n^{1-\tau_3}a_n\sum_{j\in s^*}|u_j| \geq -w_3n^{1-\tau_3}a_n\sqrt{k}. \tag{A.12}$$

By (A.10)-(A.12), we obtain

$$\begin{aligned} Q(\boldsymbol{\beta}_s) - Q(\boldsymbol{\beta}_s^*) &\leq a_n\|S(\boldsymbol{\beta}_s^*)\|_2 - \frac{c_1}{2}na_n^2 + w_3n^{1-\tau_3}a_n\sqrt{k} \\ &\leq n^{-\upsilon}\|S(\boldsymbol{\beta}_s^*)\| - \frac{c_1}{2}n^{1-2\upsilon} + w_3\sqrt{w_2}n^{1-\tau_3+\frac{1}{2}\tau_2}. \end{aligned} \tag{A.13}$$

When $\upsilon < \tau_3 - \frac{1}{2}\tau_2$, the second term dominates the third term in (A.13). Thus, for a sufficient large n and some generic constant c,

$$\begin{aligned} P\{Q(\boldsymbol{\beta}_s) \geq Q(\boldsymbol{\beta}_s^*)\} &\leq P\{\|S(\boldsymbol{\beta}_s^*)\| \geq cn^{1-\upsilon}\} \\ &\leq c\exp(-n^{1-2\upsilon-\tau_2}), \end{aligned}$$

where the last in equality is followed by the same arguments in (A.3)-(A.6). Consequently, by Bonferoni inequality and the condition $\upsilon < \frac{1}{2} - \tau_2$, we have

$$\begin{aligned} & P\{Q(\boldsymbol{\beta}_s) \geq Q(\boldsymbol{\beta}_s^*) \text{ for some } s \in \boldsymbol{S}_+^k\} \\ & \leq \sum_{s \in \boldsymbol{S}_+^k} P\{Q(\boldsymbol{\beta}_s) \geq Q(\boldsymbol{\beta}_s^*)\} \\ & \leq ckp^k \exp(-n^{1-2\upsilon-\tau_2}) \\ & \leq c \exp(\tau_2 \log n + mn^{\tau_2} \log n - n^{1-2\upsilon-\tau_2}) = o(1), \end{aligned} \tag{A.14}$$

which implies that, with probability tending to one, the local maximizer of $Q(\boldsymbol{\beta}_{\hat{s}})$ falls into a_n-neighborhood of $\boldsymbol{\beta}_{\hat{s}}^*$. The theorem is proved. ■

A.4 Proof of Theorem 2.4

Theorem 2.4 shows that, under additional requirements on the penalty function, the SMLE-PLM procedure consistently identifies the true model s^* in ultra-high dimensional situations. Similarly to the proof of Theorem 2.3, the randomness from both screening and PLM steps in the SMLE-PLM needs to be considered. We illustrate the theorem by showing that the probability of MPLE that identifies s^* from any possible $\hat{s}$ goes to one at a common rate.

Proof: Clearly, the requirements of Theorem 2.4 imply that $\tau_3 > 0.5 - \tau_2 > \tau_1 + \frac{\tau_2}{2}$. Thus, following the arguments in the proof of Theorem 2.3, there exists a local maximizer $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$ of (2.13), such that

$$\|\hat{\boldsymbol{\beta}}_\lambda(s) - \boldsymbol{\beta}_s^*\| \leq a_n = cn^{-(0.5-\tau_2-\delta)} \tag{A.15}$$

with probability tending to 1 for some generic constant c, δ and $s \in \boldsymbol{S}_+^k$.

For the convenience of presentation, we denote $\boldsymbol{\beta}_s^*$ by $\{\boldsymbol{\beta}_{s1}^*, \boldsymbol{\beta}_{s2}^*\}$ with $\boldsymbol{\beta}_{s2}^* = 0$ for any $s \in \boldsymbol{S}_+^k$. Then, for $\boldsymbol{\beta}_s = \{\boldsymbol{\beta}_{s1}, \boldsymbol{\beta}_{s2}\}$ such that $\|\boldsymbol{\beta}_s - \boldsymbol{\beta}_s^*\|_2 \leq a_n$, we have

$$Q(\boldsymbol{\beta}_{s1}, \boldsymbol{\beta}_{s2}) - Q(\boldsymbol{\beta}_{s1}^*, 0) = l(\boldsymbol{\beta}_{s1}, \boldsymbol{\beta}_{s2}) - l(\boldsymbol{\beta}_{s1}^*, 0) - n \sum_{j \in s \setminus s^*} \phi_\lambda(|\beta_j|) \tag{A.16}$$

For sufficient large n, $\boldsymbol{\beta}_s$ falls into small neighborhood of $\boldsymbol{\beta}_s^*$ such that conditions C3 and C5 become applicable. Thus, following the similar arguments in (A.3), the likelihood term in (A.16) is bounded by

$$
\begin{aligned}
& l(\boldsymbol{\beta}_{s1}, \boldsymbol{\beta}_{s2}) - l(\boldsymbol{\beta}^*_{s1}, 0) \\
& \leq \frac{\partial l(\boldsymbol{\beta}_{s1}, 0)^T}{\partial \boldsymbol{\beta}_{s2}} \boldsymbol{\beta}_{s2} - \frac{c_1}{2} n \|\boldsymbol{\beta}_{s2}\|^2 \\
& \leq \sum_{j \in s \setminus s^*} (|S_j(\boldsymbol{\beta}^*_s)| + c_5 n \|\boldsymbol{\beta}_{s1} - \boldsymbol{\beta}^*_{s1}\|) |\beta_j| \\
& \leq \|S(\boldsymbol{\beta}^*_s)\| \cdot \|\boldsymbol{\beta}_{s2}\| + c_5 n \sqrt{k} \|\boldsymbol{\beta}_{s1} - \boldsymbol{\beta}^*_{s1}\| \cdot \|\boldsymbol{\beta}_{s2}\| \qquad \text{(A.17)}
\end{aligned}
$$

Moreover, with $0 < \delta < 0.5 - \tau_1 - \tau_2$, $\boldsymbol{\beta}_{s2}$ falls into $0.5n^{-\tau_1}$ neighborhood of zero and property P4 of $\phi_\lambda(.)$ implies that

$$
\phi(|\beta_j|) = \phi^{'}(|\tilde{\beta}_j|) \text{sign}(\tilde{\beta}_j) \beta_j \geq w_4 n^{-\tau_4} |\beta_j| \qquad \text{(A.18)}
$$

for some $\tilde{\beta}_j \in (0, \beta_j)$ and $j \in s \setminus s^*$.

Consequently, with (A.17) and (A.18), we have (A.16) bounded by

$$
\begin{aligned}
& Q(\boldsymbol{\beta}_{s1}, \boldsymbol{\beta}_{s2}) - Q(\boldsymbol{\beta}^*_{s1}, 0) \\
& \leq \|S(\boldsymbol{\beta}^*_s)\| \cdot \|\boldsymbol{\beta}_{s2}\| + c_5 n \sqrt{k} \|\boldsymbol{\beta}_{s1} - \boldsymbol{\beta}^*_{s1}\| \cdot \|\boldsymbol{\beta}_{s2}\| - w_4 n^{1-\tau_4} \|\boldsymbol{\beta}_{s2}\| \\
& \leq \|S(\boldsymbol{\beta}^*_s)\| \cdot \|\boldsymbol{\beta}_{s2}\| + c_5 n^{0.5+1.5\tau_2+\delta} \|\boldsymbol{\beta}_{s2}\| - w_4 n^{1-\tau_4} \|\boldsymbol{\beta}_{s2}\| \qquad \text{(A.19)}
\end{aligned}
$$

By choosing $\delta = 0.5 \min\{0.5 - 1.5\tau_2 - \tau_4, 0.5 - \tau_1 - \tau_2\}$, the third term dominates the second term in (A.19). Thus, for a sufficiently large n and some generic constant c, we have

$$
\begin{aligned}
P\{Q(\boldsymbol{\beta}_{s1}, \boldsymbol{\beta}_{s2}) \geq Q(\boldsymbol{\beta}^*_{s1}, 0)\} & \leq P\{\|S(\boldsymbol{\beta}^*_s)\| \geq c n^{1-\tau_4}\} \\
& \leq c \exp(-n^{1-2\tau_4-\tau_2})
\end{aligned}
$$

Consequently, following the similar arguments in (A.14) with the condition $\tau_4 < 0.5 - 1.5\tau_2$, we have

$$
\begin{aligned}
& P\{Q(\boldsymbol{\beta}_{s1}, \boldsymbol{\beta}_{s2}) \geq Q(\boldsymbol{\beta}^*_{s1}, 0) \text{ for some } s \in S^k_+\} \\
& \leq ckp^k \exp(-n^{1-2\tau_4-\tau_2}) = o(1),
\end{aligned}
$$

which implies that, with probability tending to one, the local maximizer of $Q(\boldsymbol{\beta})$, say $\hat{\boldsymbol{\beta}}_\lambda(\hat{s})$, is located at $\hat{\beta}_{\lambda j}(\hat{s}) = 0$ for $j \in \hat{s} \setminus s^*$. The theorem is proved. ■

Appendix B

Supplementary Information for Chapter 3

We provide proofs of Theorems 3.1-3.2, Corollary 3.1 and the technical deviation of sample-based BIC 3.7 in this appendix. For simplicity of presentation, we use λ instead of λ_n in the proof.

B.1 Proof of Theorem 3.1

Theorem 3.1 shows that, with appropriate sampling schemes and choices of penalty functions, the PPLM consistently estimate the model coefficients and identifies the true model under the joint randomization framework. For the sampling plans where the sample mean converges at the same rate as the i.i.d. sampling, the maximizer of the penalized pseudo-likelihood necessarily converges to the census version based on the full likelihood – this further ensures the consistency of PPLM as stated in the theorem.

Proof: Let us first work on the estimation consistency of β. Let $\mathbf{u}$ be an arbitrary p-dimensional vector with $||\mathbf{u}|| = c$ and $a_n = n^{-\frac{1}{2}} + \varphi_\lambda$ for some $c > 0$. To obtain the estimation consistency, it suffices to show that as $n \to \infty$, $\hat{\boldsymbol{\beta}}_\lambda$ is in the a_n-neighborhood of $\boldsymbol{\beta}^*$.

Let $\boldsymbol{\beta} = \boldsymbol{\beta}^* + a_n\mathbf{u}$. We have

$$\begin{aligned}
&Q_n(\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta}^*) \\
&= \{l_n(\boldsymbol{\beta}^* + a_n\mathbf{u}) - l_n(\boldsymbol{\beta}^*)\} - n\sum_{j=1}^{p}[\psi_\lambda(|\boldsymbol{\beta}_j^* + a_n u_j|)] - \psi_\lambda(|\boldsymbol{\beta}_j^*|))] \\
&\leq \{l_n(\boldsymbol{\beta}^* + a_n\mathbf{u}) - l_n(\boldsymbol{\beta}^*)\} - n\sum_{j=1}^{q}[\psi_\lambda(|\boldsymbol{\beta}_j^* + a_n u_j|) - \psi_\lambda(|\boldsymbol{\beta}_j^*|)].
\end{aligned} \tag{B.1}$$

Let us work on the first term in (B.1). Define

$$H_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \mathbf{x}_i b''[\mathbf{X}_i^T\boldsymbol{\beta}]\mathbf{x}_i^T, \quad H_n(\boldsymbol{\beta}) = \sum_{i\in d} w_i\mathbf{x}_i b''[\mathbf{X}_i^T\boldsymbol{\beta}]\mathbf{x}_i^T.$$

By condition C1, when $\boldsymbol{\beta}$ is close to $\boldsymbol{\beta}^*$, each element in matrix $H_N(\boldsymbol{\beta})$, say $h_{tj} = \sum_{i=1}^{N} b''(\mathbf{x}_i\boldsymbol{\beta})x_{it}x_{ir}$ for $t, j \in \{1, \ldots, p\}$, satisfies $N^{-1}|h_{tj}|^{1+\eta} = O(1)$ with probability tending to 1. Thus, by Condition C3, we have

$$\frac{1}{n}H_n(\boldsymbol{\beta}) - \frac{1}{N}H_N(\boldsymbol{\beta}) \to_p 0. \tag{B.2}$$

Clearly, when n is large enough, $\boldsymbol{\beta}$ is in a small neighborhood of $\boldsymbol{\beta}^*$ and so (B.2) becomes applicable. Therefore, from the continuity of $I(\boldsymbol{\beta})$ at $\boldsymbol{\beta}^*$, we have

$$\begin{aligned}
l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*) &= a_n l_n'(\boldsymbol{\beta}^*)^T\mathbf{u} - \frac{1}{2}a_n^2\mathbf{u}^T H_n(\tilde{\boldsymbol{\beta}})\mathbf{u} \\
&= a_n l_n'(\boldsymbol{\beta}^*)^T\mathbf{u} - \frac{1}{2}na_n^2\mathbf{u}^T I(\boldsymbol{\beta}^*)\mathbf{u}(1 + o_p(1)) \\
&\leq a_n l_n'(\boldsymbol{\beta}^*)^T\mathbf{u} - \frac{1}{2}nc^2a_n^2M_1(1 + o_p(1)) \qquad \text{(B.3)}
\end{aligned}$$

for some $\tilde{\boldsymbol{\beta}}$ between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$, where $l_n'(\boldsymbol{\beta}^*) = \partial l_n(\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta} = \{l_{n1}'(\boldsymbol{\beta}^*), \ldots, l_{np}'(\boldsymbol{\beta}^*)\}^T$ and $l_{nj}'(\boldsymbol{\beta}^*) = \sum_{i\in d} w_i(y_i - b'(x_i\boldsymbol{\beta}^*))x_{ij}$ for $j \in \{1, \ldots, p\}$. Note that for

$j \in \{1, \ldots, p\}$,

$$\begin{aligned} E[(\{Y - b'(\mathbf{X}\boldsymbol{\beta}^*)\}X_j)^2] &= E\,\{\, E\,[\, ((Y - b'(\mathbf{X}\boldsymbol{\beta}^*))X_j)^2 |\, \mathbf{X}\,]\,\} \\ &= E\,\{b''(\mathbf{X}\boldsymbol{\beta}^*)X_j^2\,\} < \infty. \end{aligned}$$

Thus, $N^{-1}\sum_{i=1}^N (y_i - b'(x_i\boldsymbol{\beta}^*))x_{ij} = O_p(N^{-1/2})$. This and Condition C3 imply that

$$l'_{nj}(\boldsymbol{\beta}^*) = \frac{n}{N}\sum_{i=1}^N (y_i - b'(x_i\boldsymbol{\beta}^*))x_{ij} + O_p(\sqrt{n}) = O_p(\sqrt{n}). \qquad \text{(B.4)}$$

Therefore, the first term in (B.3) is $O_p(\sqrt{n}a_n) = O_p(na_n^2)$. By taking a sufficiently large c for $\|\mathbf{u}\|$, (B.3) is dominated by its second term, which implies that $P\{l_n(\boldsymbol{\beta}) \geq l_n(\boldsymbol{\beta}^*)\} \to 0$.

Now let us consider the second term in (B.1). Based on Taylor's expansion and the continuity of $\psi'_\lambda(|\beta|)$ at $\boldsymbol{\beta}_{0j}$, we have

$$\psi_\lambda(|\boldsymbol{\beta}_j^* + a_n u_j|) - \psi_\lambda(|\boldsymbol{\beta}_j^*|) = a_n u_j \psi'_\lambda(|\boldsymbol{\beta}_j^*|)\text{sign}(\beta_j^*)(1 + o(1))$$

for $j \in \{1, \ldots, q\}$. Thus, the second term in (B.1) is $O(na_n\varphi_\lambda) = O_p(na_n^2)$, which is also dominated by the second term in (B.3) for a sufficiently large c. This implies that $P\{Q_n(\boldsymbol{\beta}) \geq Q_n(\boldsymbol{\beta}^*)\} \to 0$, and therefore there exists a local maximizer of Q_n in the a_n-neighborhood of $\boldsymbol{\beta}^*$. The proof of the estimation consistency is complete.

We now examine the selection consistency. Recall that we write $\boldsymbol{\beta}^* = \{\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*\}$ with $\boldsymbol{\beta}_2^* = 0$. Hence, for any $\boldsymbol{\beta} = \{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq ca_n$ and $\boldsymbol{\beta}_2 \neq 0$, we have

$$Q_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - Q_n(\boldsymbol{\beta}_1, 0) \;=\; \{l_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - l_n(\boldsymbol{\beta}_1, 0)\} - \{n\sum_{j=q+1}^p \psi_\lambda(|\beta_j|)\} \qquad \text{(B.5)}$$

Similarly to the proof of the estimation consistency, we have

$$
\begin{aligned}
l_n(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)-l_n(\boldsymbol{\beta}_1,0) &= \frac{\partial l_n(\boldsymbol{\beta}_1,0)}{\partial \boldsymbol{\beta}_2}^T\boldsymbol{\beta}_2+\frac{1}{2}\boldsymbol{\beta}_2^T\frac{\partial l_n(\boldsymbol{\beta}_1,\tilde{\boldsymbol{\beta}}_2)}{\partial \boldsymbol{\beta}_2\partial \boldsymbol{\beta}_2^T}\boldsymbol{\beta}_2 \\
&\leq \frac{\partial l_n(\boldsymbol{\beta}_1,0)}{\partial \boldsymbol{\beta}_2}^T\boldsymbol{\beta}_2-\frac{1}{2}nM_1\|\boldsymbol{\beta}_2\|^2(1+o_p(1)) \\
&\leq \|\frac{\partial l_n(\boldsymbol{\beta}_1,0)}{\partial \boldsymbol{\beta}_2}\|\cdot\|\boldsymbol{\beta}_2\| \qquad \text{(B.6)}
\end{aligned}
$$

for some $\tilde{\boldsymbol{\beta}}_2$ between $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_2^*=0$. Also, by Taylor's expansion, it can be shown that

$$
\frac{\partial l_n(\boldsymbol{\beta}_1,0)}{\partial \beta_{2j}}=\frac{\partial l_n(\boldsymbol{\beta}^*)}{\partial \beta_{2j}}+\frac{\partial^2 l_n(\boldsymbol{\beta}^*)}{\partial \beta_{2j}\partial \boldsymbol{\beta}_1}^T(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)(1+o_p(1)) \qquad \text{(B.7)}
$$

for $j\in\{q+1,\ldots,p\}$, which implies that (B.6) is $O_p(n\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|\|\boldsymbol{\beta}_2\|+\sqrt{n}\|\boldsymbol{\beta}_2\|)$.

Note that when n is large enough, $\boldsymbol{\beta}_2$ is in a small neighborhood of 0 such that property D3 of $\psi(|\beta|)$ becomes applicable. Specifically, if $\dot{\beta}_j=0.5\beta_j$, then for some $\tilde{\beta}_j$ between β_j and $\dot{\beta}_j$, we have

$$
\begin{aligned}
\psi(|\beta_j|) &= \psi(|\dot{\beta}_j|)+\frac{1}{2}\psi'(|\tilde{\beta}_j|)\mathrm{sign}(\tilde{\beta}_j)\beta_j \\
&\geq \frac{1}{2}M_2a_n|\beta_j| \qquad \text{(B.8)}
\end{aligned}
$$

for each $j\in\{q+1,\ldots,p\}$ and an arbitrary positive constant M_2. Thus, by choosing a sufficiently large M_2, (B.5) is dominated by its second term. This implies that $P\{Q_n(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\geq Q_n(\boldsymbol{\beta}_1,0)\}\to 0$, and therefore the local maximizer of Q_n is located at some $\hat{\boldsymbol{\beta}}_{\lambda 2}=\boldsymbol{\beta}_2^*=0$. The theorem is proved. ■

B.2 Proof of Corollary 3.1

Corollary 3.1 shows that, with additional requirements on the penalty function, the PPLM is able to consistently identify the true model s and estimate their coefficients as efficiently as the MLE based on the true model. This result corresponds to the notation of oracle property in Fan and Li [2001] for the non-survey PLMs.

Proof: Clearly, the requirement for $\phi_\lambda(|\beta|)$ implies that $\varphi_\lambda \to 0$. Thus, by Theorem 1, we have $P(\hat{\boldsymbol{\beta}}_{2\lambda} = 0) \to 1$, which implies that, with probability tending to 1, the following relationship holds for $\hat{\boldsymbol{\beta}}_{1\lambda}$:

$$\frac{\partial l_n(\hat{\boldsymbol{\beta}}_{1\lambda}, 0)}{\partial \boldsymbol{\beta}_1} + \sum_{j=1}^{q} \phi'(|\hat{\beta}_{j\lambda}|) = 0. \tag{B.9}$$

Also, by Theorem 3.1, $\|\hat{\boldsymbol{\beta}}_{1\lambda} - \boldsymbol{\beta}_1^*\| = O_p(n^{-1/2} + \phi_\lambda)$, so with probability tending to 1 $\hat{\boldsymbol{\beta}}_{1\lambda}$ is in a small neighborhood of $\boldsymbol{\beta}_1^*$, which is bounded away from 0. Since for any $\beta \neq 0$ there exists $M > 0$ such that $\psi'_\lambda(|\beta|) = 0$ when $n > M$, we have

$$P(\phi'(|\hat{\beta}_{j\lambda}|) = 0) \to 1 \quad \text{for} \quad j = 1, \ldots, q.$$

This together with (B.9) implies that, with probability tending to 1, $\hat{\boldsymbol{\beta}}_{\lambda 1}$ satisfies

$$\frac{\partial l_n(\hat{\boldsymbol{\beta}}_{1\lambda}, 0)}{\partial \boldsymbol{\beta}_1} = 0,$$

which is exactly the same as the normal equation in solving the MPLE of $l_n(\boldsymbol{\beta})$ based on the true model s^*. The proof is complete. ■

B.3 Derivation of BIC (3.7)

We use the same principle as in the classical BIC (Schwarz [1978]) to obtain the sample-based BIC (3.7). Simplistically, we show that the rationale of using BIC (3.7) is to approximately maximize the pseudo-posterior (3.6) when the sample size is large. Our derivations are heuristic and we do not spell out the conditions most rigorously. The ultimate justification of (3.7) is the large sample properties of the resulting variable selection procedure, which is discussed in Theorem 3.2.

For simplicity of notation, we use s instead of s_λ in the following derivation. Let $\nu_s(\boldsymbol{\beta}_s)$ be the prior density function of regression coefficient vector $\boldsymbol{\beta}_s$ given candidate model s. A pseudo-marginal density function of the data is then given by

$$P_n(\mathbf{y}|s) = \int L_n(\mathbf{y}; \boldsymbol{\beta}_s) \nu_s(\boldsymbol{\beta}_s) d\boldsymbol{\beta}_s.$$

Consequently, we regard the following expression as the pseudo-posterior probability of model s,

$$P_n(s|\mathbf{y}) = \frac{P_n(\mathbf{y}|s)P(s)}{\sum_{s\in S} P(s)P_n(\mathbf{y}|s)},$$

where $P(s)$ denotes the prior probability of model s and S is the collection of candidate models. In the spirit of Bayesian inference, we select the model that maximizes $P_n(s|\mathbf{y})$. Since $\sum_{s\in S} P(s)P_n(\mathbf{y}|s)$ does not depend on any specific s, the highest $P_n(s|\mathbf{y})$ is achieved at the model that maximizes $P_n(\mathbf{y}|s)$ when a uniform prior $P(\cdot)$ over the model space is adopted.

To take a close look at $P_n(\mathbf{y}|s)$, let $\hat{\boldsymbol{\beta}}_s$ be the maximizer of $L_n(\mathbf{y}; \boldsymbol{\beta}_s)$ given s. Assume that $\nu_s(.)$ is smooth and does not change with the sample size n. Also, when n is large enough,

$$L_n(\mathbf{y}; \boldsymbol{\beta}_s)/L_n(\mathbf{y}; \hat{\boldsymbol{\beta}}_s) = o_p(1)$$

for $\boldsymbol{\beta}_s$ outside of a $O_p(n^{-1/2})$ neighborhood Δ of $\hat{\boldsymbol{\beta}}_s$. In addition, within $O_p(n^{-1/2})$ neighborhood of $\hat{\boldsymbol{\beta}}_s$, $\nu_s(\cdot)$ is a constant function in Δ. That is, $\nu_s(\boldsymbol{\beta}_s) \approx \nu_s(\hat{\boldsymbol{\beta}}_s)$ for $\boldsymbol{\beta}_s \in \Delta$. Thus,

$$\begin{aligned} P_n(\mathbf{y}|s) &\approx \int_{\Delta_{\hat{\boldsymbol{\beta}}_s}} L_n(\mathbf{y}; \boldsymbol{\beta}_s)\nu_s(\boldsymbol{\beta}_s)d\boldsymbol{\beta}_s \\ &\approx \nu_s(\hat{\boldsymbol{\beta}}_s) \cdot \int_{\Delta_{\hat{\boldsymbol{\beta}}_s}} L_n(\mathbf{y}; \boldsymbol{\beta}_s)d\boldsymbol{\beta}_s. \qquad \text{(B.10)} \end{aligned}$$

Let $H_n(\boldsymbol{\beta}) = -\partial^2 l_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$. Assume that $H_n(\boldsymbol{\beta})$ is continuous at $\hat{\boldsymbol{\beta}}_s$. Then, for $\boldsymbol{\beta}_s$ close to $\hat{\boldsymbol{\beta}}_s$, $l_n(\boldsymbol{\beta}_s) = \log L_n(\mathbf{y}; \boldsymbol{\beta}_s)$ is approximated by

$$l_n(\boldsymbol{\beta}_s) \approx l_n(\hat{\boldsymbol{\beta}}_s) - (1/2)(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s)^T H_n(\hat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s). \qquad \text{(B.11)}$$

Approximation (B.11) together with (B.10) implies that

$$\begin{aligned}
P_n(\mathbf{y}|s) &\approx \nu_s(\hat{\boldsymbol{\beta}}_s) \cdot \int_{\Delta_{\hat{\boldsymbol{\beta}}_s}} L_n(\mathbf{y}; \hat{\boldsymbol{\beta}}_s) \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s)^T H_n(\hat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s)\right\} d\boldsymbol{\beta}_s \\
&\approx \nu_s(\hat{\boldsymbol{\beta}}_s) \cdot L_n(\mathbf{y}; \hat{\boldsymbol{\beta}}_s) \cdot \int \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s)^T H_n(\hat{\boldsymbol{\beta}}_s)(\boldsymbol{\beta}_s - \hat{\boldsymbol{\beta}}_s)\right\} d\boldsymbol{\beta}_s \\
&= \nu_s(\hat{\boldsymbol{\beta}}_s) \cdot L_n(\mathbf{y}; \hat{\boldsymbol{\beta}}_s) \cdot (2\pi)^{\frac{\tau(s)}{2}} \cdot |H_n(\hat{\boldsymbol{\beta}}_s)|^{-\frac{1}{2}} \\
&= \nu_s(\hat{\boldsymbol{\beta}}_s) \cdot L_n(\mathbf{y}; \hat{\boldsymbol{\beta}}_s) \cdot (2\pi/n)^{\frac{\tau(s)}{2}} \cdot |n^{-1} H_n(\hat{\boldsymbol{\beta}}_s)|^{-\frac{1}{2}}, \qquad \text{(B.12)}
\end{aligned}$$

where $|.|$ denotes the determinant of a matrix and the first step is from the Laplace approximation. Consequently, we have $-2\log P_n(\mathbf{y}|s)$ approximated by

$$-2l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s)\log n + R, \qquad \text{(B.13)}$$

where

$$R = -\tau(s)\log(2\pi) + \log[\nu_s(\hat{\boldsymbol{\beta}}_s)] + \log[|n^{-1}H_n(\hat{\boldsymbol{\beta}}_s)|] = O_p(1).$$

The order is justified when

$$n^{-1}H_n(\hat{\boldsymbol{\beta}}_s) \to H(\boldsymbol{\beta}_s^*)$$

in probability for some positive definite matrix H. When the $O_p(1)$ term is ignored from (B.13), we obtain a simplified criterion

$$\mathrm{BIC}_n(s) = -2l_n(\hat{\boldsymbol{\beta}}_s) + \tau(s)\log n.$$

B.4 Proof of Theorem 3.2

Theorem 3.2 shows that the sample-based BIC score (3.7) is minimized on the true model with probability tending to one. With appropriate sampling plans, the sample-based BIC essentially reduces to the classic BIC up to a $o_p(1)$ term due to the unequal weighting, which has no effect on the consistency of the BIC-based selection.

Proof: Let $\hat{\boldsymbol{\beta}}_s$ be the maximizer of $l_n(\boldsymbol{\beta})$ based on model s. For any $\lambda \in \Omega_+ \cup \Omega_-$, we have

$$\text{BIC}_n(s_\lambda) - \text{BIC}_n(s^*) = 2\{l_n(\hat{\boldsymbol{\beta}}_{s^*}) - l_n(\hat{\boldsymbol{\beta}}_{s_\lambda})\} - \{\tau(s^*) - \tau(s_\lambda)\}\log n.$$

Thus, Theorem 3.2 is implied if

$$P\left[\{2l_n(\hat{\boldsymbol{\beta}}_{s^*}) - l_n(\hat{\boldsymbol{\beta}}_{s_\lambda})\} \le \{\tau s_* - \tau(s_\lambda)\}\log n\right] \to 0. \tag{B.14}$$

We verify (B.14) for underfitting ($\lambda \in \Omega_-$) and overfitting ($\lambda \in \Omega_+$).

Case 1: When $\lambda \in \Omega_-$, we have $s^* \not\subset s_\lambda$. Let $\check{\boldsymbol{\beta}}_{s_\lambda}$ be $\hat{\boldsymbol{\beta}}_{s_\lambda}$ augmented with zeros corresponding to $j \in \{1, \ldots, p\}$ and $j \notin s_\lambda$. We have

$$\begin{aligned} l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) - l_n(\hat{\boldsymbol{\beta}}_{s^*}) &\le l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) - l_n(\boldsymbol{\beta}^*) \\ &= l_n(\check{\boldsymbol{\beta}}_{s_\lambda}) - l_n(\boldsymbol{\beta}^*). \end{aligned}$$

Note that for $\boldsymbol{\beta}$ close to $\boldsymbol{\beta}^*$, there exists a $\tilde{\boldsymbol{\beta}}$ such that

$$\begin{aligned} l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*) &= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T l_n'(\boldsymbol{\beta}^*) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T H_n(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ &\le \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \cdot \|l_n'(\boldsymbol{\beta}^*)\| - \frac{1}{2}M_1\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 n\{1 + o_p(1)\} \end{aligned} \tag{B.15}$$

with probability tending to 1. As shown in the proof of Theorem 1, we have $\|l_n'(\boldsymbol{\beta}^*)\| = O_p(n^{1/2})$. Thus, for $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| = n^{-\frac{1}{3}}$, (B.15) is dominated by its second term which is $-(1/2)M_1 \cdot n^{\frac{1}{3}}(1 + o_p(1))$. Because of the concavity of $l_n(\boldsymbol{\beta})$,

$$P\{l_n(\boldsymbol{\beta}) - l_n(\boldsymbol{\beta}^*) \le -\frac{1}{2}M_1 \cdot n^{\frac{1}{3}}\} \to 1$$

holds for any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \ge n^{-\frac{1}{3}}$, which includes $\check{\boldsymbol{\beta}}_{s_\lambda}$ as a special case. We therefore obtain

$$l_n(\hat{\boldsymbol{\beta}}_{s^*}) - l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) \ge \frac{1}{2}M_1 \cdot n^{\frac{1}{3}}(1 + o_p(1))$$

with probability tending to 1. This implies that (B.14) holds for any $\lambda \in \Omega_-$.

Case 2: For $\lambda \in \Omega_+$, we have

$$l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) - l_n(\hat{\boldsymbol{\beta}}_{s^*}) \quad \leq \quad l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) - l_n(\boldsymbol{\beta}^*).$$

Similarly, by conditions C2 and C3, with probability tending to 1, we have

$$\begin{aligned} l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) - l_n(\boldsymbol{\beta}^*) &\leq (\hat{\boldsymbol{\beta}}_{s_\lambda} - \boldsymbol{\beta}^*)^T l_n'(\boldsymbol{\beta}^*) - \frac{1}{2}(\hat{\boldsymbol{\beta}}_{s_\lambda} - \boldsymbol{\beta}^*)^T H_n(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_{s_\lambda} - \boldsymbol{\beta}^*)\{1 + o_p(1)\} \\ &\leq (\hat{\boldsymbol{\beta}}_{s_\lambda} - \boldsymbol{\beta}^*)^T l_n'(\boldsymbol{\beta}^*) \\ &\leq n l_n'(\boldsymbol{\beta}^*)^T I^{-1}(\boldsymbol{\beta}^*) l_n'(\boldsymbol{\beta}^*)(1 + o_p(1)) \\ &\leq \frac{1}{M_1}\|l_n'(\boldsymbol{\beta}^*)\|^2 n^{-1}(1 + o_p(1)). \end{aligned}$$

Since $\|l_n'(\boldsymbol{\beta}^*)\|^2 = O_p(n)$, we have $l_n(\hat{\boldsymbol{\beta}}_{s_\lambda}) - l_n(\boldsymbol{\beta}^*) = O_p(1)$, which implies that (B.14) holds for any $\lambda \in \Omega_+$. The theorem is proved. ■

Appendix C

Supplementary Information for Chapter 4

We provide proofs of Theorems 4.1-4.2, Corollary 4.1 and Proposition 4.1 in this appendix.

C.1 Proof of Theorem 4.1

Theorem 3.1 shows that, with an appropriate scale parameter, the iteration from the ITD procedure (4.10) reduces the value of objective function Q in (4.8). The proof is similar in showing Theorem 2.2 under the GLM context.

Proof: According to (4.10), we have

$$\begin{aligned}
Q(\boldsymbol{\eta}^{(m)}) &= u^{-1}G(\boldsymbol{\eta}^{(m)}, \boldsymbol{\eta}^{(m)}) \\
&\geq u^{-1}G(\boldsymbol{\eta}^{(m+1)}, \boldsymbol{\eta}^{(m)}) \\
&= Q(\boldsymbol{\eta}^{(m+1)}) + \frac{1}{2}u^{-1}\|\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}\|^2 \\
&\quad -\sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik}[b(\theta_k^{(m+1)}) - b(\theta_k^{(m)}) - b'(\theta_k^{(m)})(\theta_k^{m+1} - \theta_k^{(m)})].
\end{aligned}$$

By Taylor's expansion, for any θ and θ_0, we have

$$b(\theta) - b(\theta_0) - b'(\theta_0)(\theta - \theta_0) = \frac{1}{2}b''(\tilde{\theta})(\theta - \theta_0)^2$$

for some $\tilde{\theta}$ between θ and θ_0. Applying this expansion, noting that $w_{ik} \in (0, 1)$, we find

$$\begin{aligned} Q(\boldsymbol{\eta}^{(m)}) &\geq Q(\boldsymbol{\eta}^{(m+1)}) + \frac{1}{2}u^{-1}\|\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}\|^2 - \frac{1}{2}n\tau_2^{(m)}\|\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}\|^2 \\ &= Q(\boldsymbol{\eta}^{(m+1)}) + \frac{1}{2}(u^{-1} - n\tau_2^{(m)}\tau_1)\|\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}\|^2 \\ &\geq Q(\boldsymbol{\eta}^{(m+1)}) \end{aligned}$$

since we have required $u^{-1} > n\tau_2^{(m)}\tau_1$. Clearly, equality holds only if $\boldsymbol{\eta}^{(m+1)} = \boldsymbol{\eta}^{(m)}$. The theorem is hence proved. ■

C.2 Proof of Corollary 4.1

Corollary 4.1 shows that the difference between two iterations in the ITD diminishes as the procedure proceeds. It serves as an important prerequisite for the convergence of ITD.

Proof: According to Theorem 4.1, we have $Q(\boldsymbol{\eta}^{(m+1)}) \leq Q(\boldsymbol{\eta}^{(m)})$. Thus, the compact assumption implies that $\boldsymbol{\eta}^{(m)}$ stays in a bounded region. Thus, $\tau_2^* = \sup_m(\tau_2^{(m)})$ is finite based on the continuity of $b''(\cdot)$. By similar arguments to those in Theorem 1, we obtain

$$Q(\boldsymbol{\eta}^{(m)}) - Q(\boldsymbol{\eta}^{(m+1)}) \geq \frac{1}{2}(u^{-1} - n\tau_2^*\tau_1)\|\boldsymbol{\eta}^{(m+1)} - \boldsymbol{\eta}^{(m)}\|^2.$$

Consequently, the asymptotically regularity of $\{\boldsymbol{\eta}^{(m)}\}$ is implied by the convergence of $Q(\boldsymbol{\eta}^{(m)})$. ■

C.3 Proof of Theorem 4.2

Theorem 4.2 shows that the ITD procedure converges to a stationary point of the objective function Q. The convergence is established based on Corollary 4.1 and the fact that the sequence of ITD has only one limiting point.

Proof: The compact condition implies that $\{\boldsymbol{\eta}^{(m)}\}$ has at least one limit point, say, $\boldsymbol{\eta}^*$. Let $\{m_t\}$ be a subsequence such that $\lim_{m\to\infty} \boldsymbol{\eta}^{(m_t)} = \boldsymbol{\eta}^*$. By Corollary 4.1, we must also have $\lim_{m\to\infty} \boldsymbol{\eta}^{(m_t+1)} = \boldsymbol{\eta}^*$.

We now show that $\boldsymbol{\eta}^*$ is also a stationary point of $Q(\boldsymbol{\eta})$. By (4.10), we have

$$\boldsymbol{\eta}^{(m_t+1)} = \arg\min_{\boldsymbol{\xi}} G(\boldsymbol{\xi}; \boldsymbol{\eta}^{(m_t)}).$$

¿From the bivariate continuity of $G(\boldsymbol{\xi}; \boldsymbol{\eta})$, letting $t \to \infty$, we get

$$\boldsymbol{\eta}^* = \arg\min_{\boldsymbol{\xi}} G(\boldsymbol{\xi}; \boldsymbol{\eta}^*).$$

That is, $\boldsymbol{\eta}^* = \{\eta_0^*, \ldots, \eta_{K-1}^*\}$ is a local/global minimum of $G(\boldsymbol{\xi}; \boldsymbol{\eta}^*)$ with respect to $\boldsymbol{\xi}$.

Therefore, we must have

$$\left.\frac{\partial G(\boldsymbol{\xi}, \boldsymbol{\eta}^*)}{\partial \xi_j}\right|_{\boldsymbol{\xi}=\boldsymbol{\eta}^*} = 0 \tag{C.1}$$

for any $j \in \{0, \ldots, K-1\}$ where $\eta_j^* \neq 0$. Denote $\theta_k^* = \sum_{j=0}^{k-1} \eta_j^*$ and $\zeta_k = \sum_{j=0}^{k-1} \xi_j$ for $k = 1, \ldots, K-1$, and write

$$G(\boldsymbol{\xi}, \boldsymbol{\eta}^*) = uQ(\boldsymbol{\xi}) + T(\boldsymbol{\xi}, \boldsymbol{\eta}^*) \tag{C.2}$$

so that

$$T(\boldsymbol{\xi}, \boldsymbol{\eta}^*) = \frac{1}{2}\sum_{j=1}^{K-1}(\xi_j - \eta_j^*)^2 - u\sum_{k=1}^{K}\sum_{i=1}^{n} w_{ik}\{b(\zeta_k) - b(\theta_k^*) - b'(\theta_k^*)(\zeta_k - \theta_k^*)\}.$$

It can be seen that

$$\left.\frac{\partial T(\boldsymbol{\xi}, \boldsymbol{\eta}^*)}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}=\boldsymbol{\eta}^*} = 0.$$

Hence, together with (C.1) and (C.2), this fact implies that $\partial Q(\boldsymbol{\eta})/\partial \eta_j$ is zero at $\boldsymbol{\eta} = \boldsymbol{\eta}^*$ for any $j \in \{0, \ldots, K-1\}$ where $\eta_j^* \neq 0$. Therefore, $\boldsymbol{\eta}^*$ is a stationary point of Q.

Suppose $\boldsymbol{\eta}^{(m)}$ does not converge but has two limiting points, say $\boldsymbol{\eta}_1^* \neq \boldsymbol{\eta}_2^*$. By what we have just proved, both are also stationary points of $Q(\cdot)$. Let $\epsilon = \|\boldsymbol{\eta}_1^* - \boldsymbol{\eta}_2^*\|$. By Corollary 4.1, the distance between successive $\eta^{(m)}$ goes to 0 as $m \to \infty$. Thus, there must be infinitely many $\eta^{(m)}$ that are at least $\epsilon/3$ from these two limiting points. The compact condition hence implies that $\boldsymbol{\eta}^{(m)}$ has at least another limiting point that is also a stationary point of $Q(\cdot)$. Let this be $\boldsymbol{\eta}_3^*$. Using the same argument, $\eta^{(m)}$ must have an additional limiting point $\boldsymbol{\eta}_4^*$, which implies that $Q(\cdot)$ has at least four stationary points. Repeatedly applying this logic implies that $Q(\cdot)$ has infinitely many stationary points. This contradicts the assumption, and therefore implies that $\boldsymbol{\eta}^{(m)}$ converges. ■

C.4 Proof of Proposition 4.1

Proposition 4.1 provides the analytic solution to the unified optimization problem (4.11) with the SCAD penalty, which serves as a build-in block for the ITD procedure (4.10). The proof follows standard procedures in mathematical analysis.

Proof: By a result from classical mathematical analysis, the solution γ^* must be a critical point of $q(\gamma)$. To avoid unnecessary complexity, we assume that $z \geq 0$ so that $\gamma^* \geq 0$. The result for $z < 0$ can be obtained by symmetry. When $\gamma > 0$, we have

$$q'(\gamma) = \gamma - z + \kappa p'_\lambda(\gamma). \tag{C.3}$$

It can be seen that $\kappa p'_\lambda(\gamma)$ is flat on the interval $(0, \lambda]$; from the point $\gamma = \lambda$ it decreases as a straight line with slope $-\kappa/(\nu - 1)$, and it remains 0 after $\gamma = \nu\lambda$ (see Fig. C.1). A nonzero critical point of $q(\cdot)$, if any, must be the intersection of $g_1(\gamma) = \kappa p'_\lambda(\gamma)$ and the straight line $g_2(\gamma) = z - \gamma$.

Suppose $0 < \kappa \leq \nu - 1$ (case 1). There is at most one such intersection: (a) when $z < \kappa\lambda$, there is no intersection so $\gamma^* = 0$ is the only critical point and the global minimum; (b) when $\kappa\lambda < z \leq (\kappa + 1)\lambda$, the unique intersection is at $\gamma^* = z - \kappa\lambda$; (c) when $(\kappa + 1)\lambda < z \leq \nu\lambda$, the unique intersection is at

$$\gamma^* = \frac{(\nu - 1)z - \kappa\nu\lambda}{\nu - \kappa - 1};$$

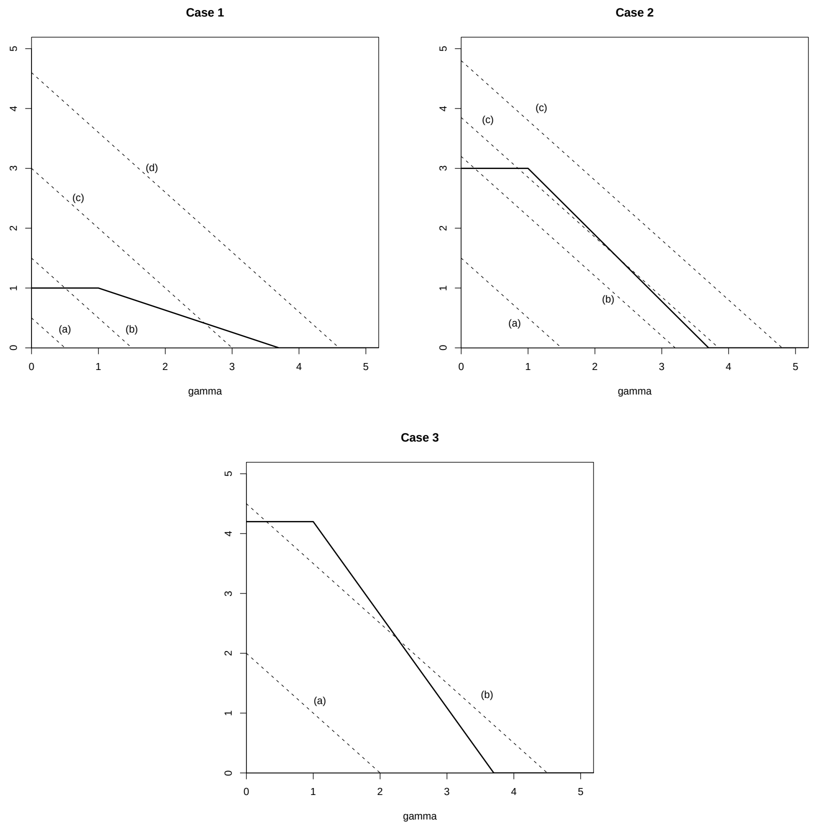

Figure C.1: Plot demonstration for Proposition 4.1 with $\nu = 3.7$, $\lambda = 1$, and $\kappa = (1, 3, 4.2)$ for cases 1–3. Solid lines represent $g_1(\gamma) = \kappa p'_\lambda(\gamma)$ and dashed lines represent $g_2(\gamma) = -\gamma + z$ for various z values.

(d) when $z > \nu\lambda$, the unique intersection is at $\gamma^* = z$. Note that cases (a) and (b) have a common expression $\gamma^* = (z - \kappa\lambda)_+$. In (b), (c), and (d), the nonzero critical point is found to be the global minimum. Thus, we have obtained the result for the case where $\kappa > \nu - 1$ when $z > 0$.

Suppose $\nu - 1 < \kappa \leq \nu$ (case 2). In this case, $\kappa p'_\lambda(\gamma)$ decreases with a slope smaller than -1 from $\gamma = \lambda$ until it reaches 0. This creates the possibility of up to four intersections between $\kappa p'_\lambda(\gamma)$ and $g(\gamma)$. Yet we find that the global minimum is at the largest critical point: (a) when $z < \kappa\lambda$, there is no intersection and we find $\gamma^* = 0$; (b) when $\kappa\lambda \leq z < \nu\lambda$, the two curves intersect at $\gamma^* = z - \kappa\lambda$; (c) when $z \geq \nu\lambda$, there can be many intersections but the global minimum is given by $\gamma^* = z$.

Suppose $\nu < \kappa$ (case 3). The derivation is the simplest in this case: (a) when $z < \nu\lambda$, there is no intersection and we find $\gamma^* = 0$; (b) when $\nu\lambda \leq z$, the largest intersection is $\gamma^* = z$, which is the global minimum.

We have considered all possible cases. For $z < 0$, the expression of γ^* must be adjusted accordingly. ■