TOPICS IN HIGH-DIMENSIONAL INFERENCE

BY WENHUA JIANG

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Statistics and Biostatistics

Written under the direction of
Professor Cun-Hui Zhang
and approved by

____

____

____

____

New Brunswick, New Jersey
October, 2009

© 2009

Wenhua Jiang

ALL RIGHTS RESERVED

ABSTRACT OF THE DISSERTATION

Topics in High-dimensional Inference

by Wenhua Jiang

Dissertation Director: Professor Cun-Hui Zhang

This thesis concerns three connected problems in high-dimensional inference: compound estimation of normal means, nonparametric regression and penalization method for variable selection.

In the first part of the thesis, we propose a general maximum likelihood empirical Bayes (GMLEB) method for the compound estimation of normal means. We prove that under mild moment conditions on the unknown means, the GMLEB enjoys the adaptive ration optimality and adaptive minimaxity. Simulation experiments demonstrate that the GMLEB outperforms the James-Stein and several state-of-the-art threshold estimators in a wide range of settings.

In the second part, we explore the GMLEB wavelet method for nonparametric regression. We show that the estimator is adaptive minimax in all Besov balls. Simulation experiments on the standard test functions demonstrate that the GMLEB outperforms several threshold estimators with moderate and large samples. Applications to high-throughput screening (HTS) data are used to show the excellent performance of the approach.

In the third part, we develop a generalized penalized linear unbiased selection (GPLUS) algorithm to compute the solution paths of concave-penalized negative log-likelihood for generalized linear model. We implement the smoothly clipped absolute deviation (SCAD) and minimax concave (MC) penalties in our simulation study to demonstrate the feasibility of the proposed algorithm and their superior selection accuracy compared with the ℓ_1 penalty.

Acknowledgements

I am deeply indebted to my advisor, Professor Cun-Hui Zhang for his tremendous support, invaluable guidance, and constant encouragement. During the ups and downs in the past years, Professor Zhang provided me generous help on my study and life. No word can express my gratitude. His devotion to intellect will always be a source of inspiration for me.

I wish to thank the other members of my dissertation committee, Professor Adi Ben-Israel, Professor Kesar Singh and Professor William Strawderman for their helpful comments on the manuscript. I would like to thank Professor Minge Xie for his encouragement. I want to thank the Department of Statistics and Biostatistics of Rutgers University, the faculty and staff in the department for their support. Especially I want to thank our graduate director Professor John Kolassa, who has always been very supportive. Special thanks goes to Doctor Donghui Zhang for her direction and support during my work in Sanofi-Aventis.

My thanks also go to my friends Dong Dai, Ye Li, Ming Shi, Jue Wang, Minya Xu, Fei Ye, Biao Yin and Juan Zhang. I deeply appreciate their friendship and help.

At last, I want to thank my parents for their love and support.

Dedication

To My Dear Parents

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

This thesis concerns three connected problems in high-dimensional inference: compound estimation of normal means, nonparametric regression and penalization method for variable selection.

The first problem, known as the compound estimation of normal means, has been considered as the canonical model or motivating example in the developments of empirical Bayes, admissibility, adaptive nonparametric regression, variable selection, multiple testing and many other areas in statistics. Let X_i be independent observations with $X_i \sim N(\theta_i, 1)$, $i = 1, \ldots, n$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is an unknown deterministic signal vector. Our problem is to estimate $\boldsymbol{\theta}$ under the compound loss $L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} (\widehat{\theta}_i - \theta_i)^2$. There are three main approaches in the compound estimation of normal means: the general empirical Bayes (EB) [59, 62], the James-Stein estimator [44, 64], and the threshold method [1, 5, 22, 23, 47]. Among the three approaches, the later two work well when the empirical distribution of the unknown means is approximately normal or very sparse, respectively. The general EB is greedier since it assumes essentially no knowledge about the unknown means but still aims to attain the performance of the oracle separable estimator based on the knowledge of the empirical distribution of the unknowns. Thus, the heart of the question is whether the gain by aiming at the smaller general EB benchmark risk is large enough to offset the additional cost of the nonparametric estimation.

We propose a general maximum likelihood EB (GMLEB) in which we first estimate the empirical distribution of the unknown means by the generalized maximum likelihood estimator (MLE) [49] and then plug the estimator into the oracle general EB rule. Our results affirm that by aiming at the minimum risk of all separable estimators, the greedier general EB approach realizes significant risk reduction over linear and threshold methods for a wide range of unknown signal vectors for moderate and large samples, and this is especially so for the GMLEB. We prove that the risk of the GMLEB estimator is within an infinitesimal fraction of the ideal Bayes risk when this risk of greater order than $(\log n)^5/n$ depending on the magnitude of the weak ℓ_p norm of the unknown means, $0 < p \leq \infty$. Such adaptive ratio optimality is obtained through an oracle inequality which provides a uniform upper bound of the regret using the GMLEB estimator. Moreover, we use this oracle inequality to prove the adaptive minimaxity of the GMLEB estimator over a broad collection of ℓ_p balls. We demonstrate the superb risk performance of the GMLEB for moderate samples through simulation experiments, and provide an EM algorithm for the computation of the GMLEB.

The second problem is nonparametric regression which is a typical example where the compound estimation of normal means can be directly applied. We have $N = 2^J$ noisy samples Y_i of a function f, $Y_i = f(t_i) + e_i$, $i = 1, \ldots, N$ where $t_i = i/N$ and e_i are independent $N(0, \sigma^2)$ random variables. Our goal is to recover the unknown function f. In detail, we measure the performance of an estimate $\widehat{f}$ in term of squared loss at the sample points by the risk $R(\widehat{f}, f) = N^{-1}E\|\widehat{f} - f\|^2 = N^{-1}\sum_{i=1}^{N} E(\widehat{f}(t_i) - f(t_i))^2$.

We propose the GMLEB wavelet method to the nonparametric regression problem. Our method proceeds by taking the discrete wavelet transform of the data Y_i , processing the resulting coefficients to remove noise by the GMLEB method, and then transforming back to obtain the estimate. We provide an oracle inequality, that is, an upper bound for the estimation regret for the adaptation to the ideal risk. Moreover, we show that the worst behavior of our estimation method when the function f is constrained to lie in a Besov space simultaneously attains the best possible minimax risk over a wide range of Besov spaces. This adaptive minimaxity implies the adaptation to spatial inhomogeneity of the unknown function. We conduct an extensive Monte Carlo simulation study of the performance of our estimator using four standard test functions. It turns out that for moderate and large samples, our procedure outperforms other threshold estimators and improves over a general EB method based on Fourier smoothing kernel [74].

The third problem concerns penalization methods for variable selection. In the wavelet thresholding approach to the standard nonparametric regression model, the discrete wavelet transform (DWT) proceeds by $\boldsymbol{y} = N^{-1/2}\mathcal{W}\boldsymbol{Y}$ where $\mathcal{W}$, called the finite wavelet transformation matrix, is a real N by N orthonormal matrix and $\boldsymbol{y}$ is the vector of the discrete wavelet coefficients. The distribution of $\boldsymbol{y}$ is $N(\boldsymbol{\beta}, \epsilon^2 \boldsymbol{I}_N)$ with unknown $\boldsymbol{\beta}$ where $\epsilon^2 = \sigma^2/N$. The wavelet thresholding amounts to estimate $\boldsymbol{\beta}$ based on observations $\boldsymbol{Y}$ using the linear model $\boldsymbol{Y} = N^{1/2}\mathcal{W}^T\boldsymbol{\beta} + \sigma\boldsymbol{z}$ where $\boldsymbol{z} \sim N(0, \boldsymbol{I}_N)$. This is a special case in variable selection since the design is orthogonal. General variable selection is more complicated and challenging since there are dependencies among variables.

In linear regression, a number of concave penalized least squares methods have been shown to possess selection consistency and oracle efficiency properties under much weaker conditions than the ℓ_1 penalized methods do [34, 79, 85]. However, minimization of a concave penalized general loss function is still a computationally challenging problem. A penalized linear unbiased selection (PLUS) algorithm was recently proposed for the computation of a solution path of local minimizers of concave penalized least squares [78]. The main idea of the PLUS is to compute possibly multiple local minimizers at individual penalty levels by continuously tracing the minimizers at different penalty levels. We develop a generalized PLUS (GPLUS) algorithm to compute the solution paths of concave-penalized negative log-likelihood. We use end-to-end short linear segments to approximate the nonlinear paths of generalized linear models. We implement the smoothly clipped absolute deviation (SCAD) [34] and minimax concave (MC) [78] penalties in our simulation study to demonstrate the feasibility of the proposed algorithm and their superior selection accuracy compared with the ℓ_1 penalty.

Chapter 2
General Empirical Bayes Estimation of Normal Means

2.1 Introduction

This chapter concerns the estimation of a vector with iid normal errors under the average squared loss. Let X_i be independent statistics with

$$X_i \sim \varphi(x - \theta_i) \sim N(\theta_i, 1), \quad i = 1, \ldots, n, \tag{2.1}$$

under a probability measure $P_{n,\boldsymbol{\theta}}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is an known signal vector. Our problem is to estimate $\boldsymbol{\theta}$ under the compound loss

$$L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\theta}_i - \theta_i)^2 \tag{2.2}$$

for any given estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$. Throughout this chapter, the unknown means θ_i are assumed to be deterministic as in the standard compound decision theory [59]. To avoid confusion, the Greek θ is used only with boldface as a deterministic mean vector $\boldsymbol{\theta}$ in $\mathbb{R}^n$ or with subscripts as elements of $\boldsymbol{\theta}$.

The problem, known as the compound estimation of normal means, has been considered as the canonical model or motivating example in the developments of empirical Bayes, admissibility, adaptive nonparametric regression, variable selection, multiple testing and many other areas in statistics. It also carries significant practical relevance in statistical applications since the observed data are often understood, represented or summarized as the sum of a signal vector and the white noise.

There are three main approaches in the compound estimation of normal means. The first one is the general empirical Bayes (EB) [59, 62], which assumes essentially no knowledge about the unknown means but still aims to attain the performance of the oracle separable estimator based on the knowledge of the empirical distribution of the unknowns. Here a separable estimator is one that uses a fixed deterministic function of the i-th observation to estimate the i-th mean. This greedy approach, also called nonparametric EB [54], was proposed the earliest among the three, but it is also the least understood, in spite of [60, 61, 62, 74, 75, 76]. Efron [28] attributed this situation to the lack of applications with many unknowns before the information era and pointed out that "current scientific trends favor a greatly increased role for empirical Bayes methods" due to the prevalence of large, high-dimensional data and the rapid rise of computing power. The methodological and theoretical challenge, which we focus on in this chapter, is to find the "best" general EB estimators and sort out the type and size of problems suitable for them.

The second approach, conceived with the celebrated Stein's proof of the inadmissibility of the optimal unbiased estimator and the introduction of the James-Stein estimator [44, 64], is best understood through its parametric or linear EB interpretations [30, 31, 54]. The James-Stein estimator is minimax over the entire space of the unknown mean vector and well approximates the optimal linear separable estimator based on the oracular knowledge of the first two empirical moments of the unknown means. Thus, it achieves the general EB optimality when the empirical distribution of the unknown means are approximately normal. However, the James-Stein estimator does not perform well by design compared with the general EB when the minimum risk of linear separable estimators is far different from that of all separable estimators [74]. Still, what is the cost of being greedy with the general EB when the empirical distribution of the unknown means is indeed approximately normal?

The third approach focuses on unknown mean vectors which are sparse in the sense of having many (near) zeros. Such sparse vectors can be treated as members of small ℓ_p balls with $p < 2$. Examples include the estimation of functions with unknown discontinuity or inhomogeneous smoothness across different parts of a domain in nonparametric regression [22]. For sparse means, the James-Stein or the linear estimators could perform much worse than threshold estimators [21]. Many threshold methods have been proposed and proved to possess (near) optimality properties for sparse signals, including the universal [22], SURE [23], FDR [1], the generalized C_p [5] and the parametric EB posterior median (EBThresh) [47]. These estimators can be viewed as approximations of the optimal candidate in certain families of separable threshold estimators, so that they do not perform well by design compared with the general EB when the minimum risk of separable threshold estimators is far different from that of all separable estimators [76]. Again, what is the cost of being greedy with the general EB when the unknown means are indeed very sparse?

Since general EB methods have to spend more "degrees of freedom" for nonparametric estimation of its oracle rule, compared with linear and threshold methods, the heart of the question is whether the gain by aiming at the smaller general EB benchmark risk is large enough to offset the additional cost of the nonparametric estimation.

We propose a general maximum likelihood EB (GMLEB) in which we first estimate the empirical distribution of the unknown means by the generalized maximum likelihood estimator (MLE) [49] and then plug the estimator into the oracle general EB rule. In other words, we treat the unknown means as iid variables with a completely unknown common "prior" distribution (for the purpose of deriving the GMLEB, whether the unknowns are actually deterministic or random), estimate the nominal prior with the generalized MLE, and then use the Bayes rule for the estimated prior. The basic idea was discussed in the last paragraph of [59] as a general way of deriving solutions to compound decision problems, although the notion of MLE was vague at that time without a parametric model and not much has been done since then about using the generalized MLE to estimate the nominal prior in compound estimation.

Our results affirm that by aiming at the minimum risk of all separable estimators, the greedier general EB approach realizes significant risk reduction over linear and threshold methods for a wide range of unknown signal vectors for moderate and large samples, and this is especially so for the GMLEB. We prove that the risk of the GMLEB estimator is within an infinitesimal fraction of the general EB benchmark when the risk is of the order $n^{-1}(\log n)^5$ or greater depending on the magnitude of the weak ℓ_p norm of the unknown means, $0 < p \leq \infty$. Such adaptive ratio optimality is obtained through a general oracle inequality which also implies the adaptive minimaxity of the GMLEB over a broad collection of regular and weak ℓ_p balls. This adaptive minimaxity result unifies and improves upon the adaptive minimaxity of threshold estimators for sparse means [1, 23, 47] and the Fourier general EB estimators for moderately sparse and dense means [74]. We demonstrate the superb risk performance of the GMLEB for moderate samples through simulation experiments, and describe algorithms to show its computational feasibility.

This chapter is organized as follows. In Section 2.2, we introduce the general EB method. In Section 2.3, we introduce the GMLEB method and its computation. In Section 2.4, we provide upper bounds for the regret of a regularized Bayes rule using a predetermined and possibly misspecified prior and prove an oracle inequality for the GMLEB, compared with the general EB benchmark risk. The consequences of this oracle inequality, including statements of our adaptive ratio optimality and adaptive minimaxity results in full strength, are also discussed in Section 2.4. In Section 2.5, we introduce a regularized Fourier general EB (RF-GEB) estimator and provide an oracle inequality for it. In Section 2.6, we present some simulation results. Section 2.7 contains some discussions. Mathematical proofs of theorems. propositions and lemmas are given either right after their statements or in Section 2.8.

2.2 The Empirical Bayes Method

Throughout the chapter, boldface letters denote vectors and matrices, for example, $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ denotes the standard normal density, $\widetilde{L}(y) = \sqrt{-\log(2\pi y^2)}$ denotes the inverse of $y = \varphi(x)$ for positive x and y, $x \vee y = \max(x, y)$, $x \wedge y = \min(x, y)$, $x_+ = x \vee 0$ and $a_n \asymp b_n$ means $0 < a_n/b_n + b_n/a_n = O(1)$. In a number of instances, $\log(x)$ should be viewed as $\log(x \vee e)$. Univariate functions are applied to vectors per component. Thus, an estimator of $\boldsymbol{\theta}$ is separable if it is of the form $\widehat{\boldsymbol{\theta}} = t(\boldsymbol{X}) = (t(X_1), \ldots, t(X_n))$ with a predetermined Borel function $t(\cdot)$. In the vector notation, it is convenient to state (2.1) as $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ with $\boldsymbol{I}_n$ being the identity matrix in $\mathbb{R}^n$.

The compound estimation of a vector of deterministic normal means is closely related to the Bayes estimation of a single random mean. In this Bayes problem, we estimate a univariate random parameter ξ based on a univariate Y such that

$$Y|\xi \sim N(\xi, 1), \quad \xi \sim G, \quad \text{under } P_G. \tag{2.3}$$

The prior distribution $G = G_n$ which naturally matches the unknown means $\{\theta_i, i \le n\}$ in (2.1) is the empirical distribution

$$G_n(u) = G_{n,\boldsymbol{\theta}}(u) = \frac{1}{n}\sum_{i=1}^{n} I\{\theta_i \le u\}. \tag{2.4}$$

Here and in the sequel, subscripts ${}_{n,\boldsymbol{\theta}}$ indicate dependence of distribution or probability upon n and the unknown deterministic vector $\boldsymbol{\theta}$.

The *fundamental theorem of compound decisions* [59] in the context of the ℓ_2 loss asserts that the compound risk of a separable rule $\widehat{\boldsymbol{\theta}} = t(\boldsymbol{X})$ under the probability $P_{n,\boldsymbol{\theta}}$ in the multivariate model (2.1) is identical to the MSE of the same rule $\widehat{\xi} = t(Y)$ under the prior (2.4) in the univariate model (2.3):

$$E_{n,\boldsymbol{\theta}} L_n(t(\boldsymbol{X}), \boldsymbol{\theta}) = E_{G_n}(t(Y) - \xi)^2. \tag{2.5}$$

For any true or nominal priors G, denote the Bayes rule as

$$t_G^* = \arg\min_t E_G(t(Y) - \xi)^2 = \frac{\int u\varphi(Y-u)G(du)}{\int \varphi(Y-u)G(du)}, \tag{2.6}$$

and the minimum Bayes risk as

$$R^*(G) = E_G(t^*_G(Y) - \xi)^2, \tag{2.7}$$

where the minimum is taken over all Borel functions. It follows from (2.5) that among all separable rules, the compound risk is minimized by the Bayes rule with prior (2.4), resulting in the general EB benchmark

$$R^*(G_n) = E_{n,\boldsymbol{\theta}} L_n(t^*_{G_n}(\boldsymbol{X}), \boldsymbol{\theta}) = \min_{t(\cdot)} E_{n,\boldsymbol{\theta}} L_n(t(\boldsymbol{X}), \boldsymbol{\theta}). \tag{2.8}$$

The general EB approach seeks procedures which approximate the Bayes rule $t^*_{G_n}(\boldsymbol{X})$ or approximately achieve the risk benchmark $R^*(G_n)$ in (2.8).

Given a class of functions $\mathscr{D}$, the aim of the restricted EB is to attain

$$R_{\mathscr{D}}(G_n) = \inf_{t \in \mathscr{D}} E_{n,\boldsymbol{\theta}} L_n(t(\boldsymbol{X}), \boldsymbol{\theta}) = \inf_{t \in \mathscr{D}} E_{G_n}(t(Y) - \xi)^2 \tag{2.9}$$

approximately. This provides EB interpretations for all the adaptive methods discussed in the introduction, with $\mathscr{D}$ being the classes of all linear functions for the James-Stein estimator, all soft threshold functions for the SURE [23], and all hard threshold functions for the generalized C_p [5] or the FDR [1]. For the EBThresh [47], $\mathscr{D}$ is the class of all posterior median functions $t(y) = \text{median}(\xi | Y = y)$ under the probability P_G in (2.3) for priors of the form

$$G(u) = \omega_0 I(0 \leq u) + (1 - \omega_0) G_0(u/\tau), \tag{2.10}$$

where w_0 and τ are free and G_0 is given.

Compared with linear and threshold methods, the general EB approach is greedier since it aims at the smaller benchmark risk: $R^*(G_n) \leq R_{\mathscr{D}}(G_n)$ for all $\mathscr{D}$. This could backfire when the regret

$$r_{n,\boldsymbol{\theta}}(\widehat{t}_n) = E_{n,\boldsymbol{\theta}} L_n(\widehat{t}_n(\boldsymbol{X}), \boldsymbol{\theta}) - R^*(G_n) \tag{2.11}$$

of using an estimator $\widehat{t}_n(\cdot)$ of the general EB oracle rule $t^*_{G_n}(\cdot)$ is greater than the difference $R_{\mathscr{D}}(G_n) - R^*(G_n)$ in benchmark, but our simulation and oracle inequalities prove that $r_{n,\boldsymbol{\theta}}(\widehat{t}_n) = o(1) R^*(G_n)$ uniformly for a wide range of the unknown vector $\boldsymbol{\theta}$ and moderate/large samples.

Zhang [74] proposed a general EB method based on a Fourier infinite order smoothing kernel. The Fourier general EB estimator is asymptotically minimax over the entire parameter space and approximately reaches the general EB benchmark (2.8) uniformly for dense and moderately sparse signals, provided that the oracle Bayes risk is of the order $n^{-1/2}(\log n)^{3/2}$ or greater [74]. Hybrid general EB estimators have been developed [76] to combine the features and optimality properties of the Fourier general EB and threshold estimators. Still, the performance of general EB methods is sometimes perceived as uncertain in moderate samples [47]. Indeed, the Fourier general EB requires selection of certain tuning parameters and its proven theoretical properties are not completely satisfying. This motivates our investigation.

2.3 The General Maximum Likelihood Empirical Bayes Method

2.3.1 The GMLEB method

The GMLEB method replaces the unknown prior G_n of the oracle rule $t^*_{G_n}$ by its generalized MLE [49]

$$\widehat{G}_n = \widehat{G}(\cdot; \boldsymbol{X}) = \arg\max_{G \in \mathscr{G}} \prod_{i=1}^{n} f_G(X_i), \tag{2.12}$$

where $\mathscr{G}$ is the family of all distribution function and f_G is the density

$$f_G(x) = \int \varphi(x-u) G(du) \tag{2.13}$$

of the normal location mixture by distribution G.

The estimator (2.12) is called the generalized MLE since the likelihood is used only as a vehicle to generate the estimator. The G here is used only as a nominal prior. In our adaptive ratio and minimax optimality theorems and oracle inequality, the GMLEB is evaluated under the measures $P_{n,\boldsymbol{\theta}}$ in (2.1) where the unknowns θ_i are assumed to be deterministic parameters.

Since (2.12) is typically solved by iterative algorithms, we allow approximate solutions to be used. For definiteness and notation simplicity, the generalized MLE in the sequel is any solution of

$$\widehat{G}_n \in \mathscr{G}, \quad \prod_{i=1}^n f_{\widehat{G}_n}(X_i) \geq q_n \sup_{G\in\mathscr{G}} \prod_{i=1}^n f_G(X_i), \tag{2.14}$$

with $q_n = (e\sqrt{2\pi}/n^2) \wedge 1$, although the theoretical results in this chapter all hold verbatim for less stringent (2.14) with $0 \leq \log(1/q_n) \leq c_0(\log n)$ for any fixed constant c_0. Formally, the GMLEB estimator is defined as

$$\widehat{\theta}_i = t^*_{\widehat{G}_n}(X_i), \quad i = 1, \ldots, n, \tag{2.15}$$

where t^*_G is the Bayes rule in (3.16) and $\widehat{G}_n$ is any approximate generalized MLE (2.14) for the nominal prior (2.4). Clearly, the GMLEB estimator (2.15) is completely nonparametric and does not require any restriction, regularization, bandwidth selection or other forms of tuning.

The GMLEB is location equivariant in the sense that

$$t^*_{\widehat{G}_n(\cdot;\boldsymbol{X}+c\boldsymbol{e})}(\boldsymbol{X}+c\boldsymbol{e}) = t^*_{\widehat{G}_n(\cdot;\boldsymbol{X})}(\boldsymbol{X}) + c\boldsymbol{e} \tag{2.16}$$

for all real c, where $\boldsymbol{e} = (1, \ldots, 1) \in \mathbb{R}^n$. This is due to the location equivariance of the generalized MLE: $\widehat{G}_n(x; \boldsymbol{X} + c\boldsymbol{e}) = \widehat{G}_n(x - c; \boldsymbol{X})$. Compared with the Fourier general EB estimators [74, 76], the GMLEB (2.15) is more appealing since the function $t^*_{\widehat{G}_n}(x)$ of x enjoys all analytical properties of Bayes rules: monotonicity, infinite differentiability and more. However, the GMLEB is much harder to analyze than the Fourier general EB.

2.3.2 Computation of the GMLEB

It follows from the Carathéodory's theorem [17] that there exists a discrete solution of (2.12) with no more than $n + 1$ support points. A discrete approximate generalized MLE $\widehat{G}_n$ with m support points can be written as

$$\widehat{G}_n = \sum_{j=1}^m \widehat{w}_j \delta_{u_j}, \quad \widehat{w}_j \geq 0, \quad \sum_{j=1}^m \widehat{w}_j = 1, \tag{2.17}$$

where δ_u is the probability distribution giving its entire mass to u. Given (2.17), the GMLEB estimator can be easily computed as

$$\widehat{\theta}_i = t^*_{\widehat{G}_n}(X_i) = \frac{\sum_{j=1}^m u_j \varphi(X_i - u_j)\widehat{w}_j}{\sum_{j=1}^m \varphi(X_i - u_j)\widehat{w}_j}, \tag{2.18}$$

since $t^*_G(x)$ is the conditional expectation as in (3.16).

Since the generalized MLE $\widehat{G}_n$ is completely nonparametric, the support points $\{u_j, j \leq m\}$ and weights $\{\widehat{w}_j, j \leq m\}$ in (2.17) are selected or computed solely to maximize the likelihood in (2.12). There are quite a few possible algorithms for solving (2.14), but all depend on iterative approximations. Due to the monotonicity of $\varphi(t)$ in t^2, the generalized MLE (2.12) puts all its mass in the interval $I_0 = [\min_{1\leq i\leq n} X_i, \max_{1\leq i\leq n} X_i]$. Given fine grid $\{u_j\}$ in I_0, the EM-algorithm [20, 71]

$$\widehat{w}_j^{(k)} = \frac{1}{n}\sum_{i=1}^n \frac{\widehat{w}_j^{(k-1)}\varphi(X_i - u_j)}{\sum_{\ell=1}^m \widehat{w}_\ell^{(k-1)}\varphi(X_i - u_\ell)} \tag{2.19}$$

optimizes the weights $\{\widehat{w}_j\}$. In Subsection 2.7.2, we provide a conservative statistical criterion on $\{u_j\}$ and an EM-stopping rule to guarantee (2.14).

We took a simple approach in our simulation experiments. Given $\{X_i, 1 \leq i \leq n\}$ and with $X_0 = 0$, we chose the grid points $\{u_j\}$ as a set of multipliers of $\epsilon = \max_{0\leq i<j\leq n} |X_i - X_j|/999$ with $u_j = u_{j-1} + \epsilon$ and the range

$$-j_0\epsilon = u_1 - \epsilon < \min_{0\leq i\leq n} X_i \leq u_1, \quad u_m = (m - j_0)\epsilon \leq \max_{0\leq i\leq n} X_i < u_m + \epsilon$$

with an integer $j_0 \in [1, m]$. This ensures $u_{j_0} = 0$ as a grid point and $999 \leq m \leq 1000$. We ran 100 EM-iterations (2.19) in our simulations. We have tried to optimize both the support points $\{u_j\}$ and weights $\{\widehat{w}_j\}$ in the EM-algorithm, but gained limited improvements.

The GMLEB estimator (2.18) depends slightly on the initialization of the EM-algorithm due to the non-uniqueness of the GMLEB estimator and the fixed number of EM-iterations in our implementation. Since the generalized MLE (2.12) is unique only up to the values of $\{f_{\widehat{G}_n}(X_i), i \leq n\}$, different EM-initializations lead to different versions of $\widehat{G}_n$, which then result in different values of $t^*_{\widehat{G}_n}(X_i)$

in (2.18). This non-uniqueness persists even when we run infinitely many EM-iterations. Nevertheless, our theoretical results hold for all versions of the GMLEB.

We consider two options in our simulation experiments. The first option initializes the weights with the uniform distribution $\widehat{w}_j = 1/m$. The second option takes into consideration of the possible sparsity of the signal by putting a good starting mass at $u_{j_0} = 0$:

$$\widehat{w}_{j_0} = \widehat{\omega}_0, \quad \widehat{w}_j = \frac{1 - \widehat{\omega}_0}{m-1}, \quad j \neq j_0. \tag{2.20}$$

We estimate the proportion of zeros within the n means by a Fourier method,

$$\widehat{\omega}_0 = \frac{1}{n}\sum_{j=1}^{n} \psi(X_j; h_n), \quad \psi(z;h) = \int h\psi_0(ht)e^{t^2/2}\cos(zt)dt,$$

as in [66, 67], where ψ_0 is a density function with support $[-1,1]$ and $h_n = \{\kappa(\log n)\}^{-1/2}$ is the bandwidth, $\kappa \leq 1$. In our simulation experiments, the uniform $[-1,1]$ density is used as ψ_0 and $\kappa = 1/2$. To distinguish the two options of initializing the EM-algorithm, we reserve the name GMLEB for the uniform initialization and call (sparse-) S-GMLEB the estimator with the initialization (2.20) when we report simulation results.

2.4 Theoretical Properties of the GMLEB

2.4.1 A regularized Bayes estimator with a misspecified prior

In this subsection, we consider a fixed probability P_{G_0} under which

$$Y|\xi \sim N(\xi, 1), \quad \xi \sim G_0. \tag{2.21}$$

Recall [9, 60] that for the estimation of a normal mean, the Bayes rule (3.16) and its risk (2.7) can be expressed in terms of the mixture density $f_G(x)$ as

$$t^*_G(x) = x + \frac{f'_G(x)}{f_G(x)}, \quad R^*(G) = 1 - \int \left(\frac{f'_G}{f_G}\right)^2 f_G, \tag{2.22}$$

in the model (2.3), where $f_G(x) = \int \varphi(x-u)G(du)$ is as in (2.13).

Suppose the true prior G_0 is unknown but a deterministic approximation of it, say G, is available. The Bayes formula (2.22) could still be used, but we may want to avoid dividing by a near-zero quantity. This leads to the following regularized Bayes estimator:

$$t_G^*(x;\rho) = x + \frac{f_G'(x)}{f_G(x) \vee \rho}. \tag{2.23}$$

For $\rho = 0$, $t_G^*(x;0) = t_G^*(x)$ is the Bayes estimator for the prior G. For $\rho = \infty$, $t_G^*(x;\infty) = x$ gives the MLE of ξ which requires no knowledge of the prior. The following proposition, describes some analytical properties of the regularized Bayes estimator.

Proposition 2.1. *Let $\widetilde{L}(y) = \sqrt{-\log(2\pi y^2)}$, $y \geq 0$, be the inverse function of $y = \varphi(x)$. Then, the value of the regularized Bayes estimator $t_G^*(x;\rho)$ in (2.23) is always between those of the Bayes estimator $t_G^*(x)$ in (3.16) and the MLE $t_G^*(x;\infty) = x$. Moreover, for all real x*

$$\begin{cases} \big|x - t_G^*(x;\rho)\big| = \frac{|f_G'(x)|}{f_G(x) \vee \rho} \leq \widetilde{L}(\rho), & \textit{if } 0 < \rho < (2\pi e)^{-1/2}, \\ 0 \leq (\partial/\partial x) t_G^*(x;\rho) \leq \widetilde{L}^2(\rho), & \textit{if } 0 < \rho < (2\pi e^3)^{-1/2}. \end{cases} \tag{2.24}$$

Remark 2.1. *In [74], a slightly different inequality*

$$\Big(\frac{f_G'(x)}{f_G(x)}\Big)^2 \frac{f_G(x)}{f_G(x) \vee \rho} \leq \widetilde{L}^2(\rho), \quad 0 \leq \rho < (2\pi e^2)^{-1/2}, \tag{2.25}$$

was used to derive oracle inequalities for Fourier general EB estimators. The extension to the derivative of $t_G^*(x;\rho)$ here is needed for the application of the Gaussian isoperimetric inequality in Proposition 2.4.

The next theorem provides oracle inequalities which bound the regret of using (2.23) due to the lack of the knowledge of the true G_0. Let

$$d(f,g) = \Big(\int (f^{1/2} - g^{1/2})^2\Big)^{1/2} \tag{2.26}$$

denote the Hellinger distance. The upper bounds asserts that the regret is no greater than square of the Hellinger distance between the mixture densities f_G and f_{G_0} up to certain logarithmic factors.

Theorem 2.1. *Suppose (2.21) holds under P_{G_0}. Let $t^*_G(x;\rho)$ be the regularized Bayes rule in (2.23) with $0<\rho\le(2\pi e^2)^{-1/2}$. Let f_G be as in (2.13).*

(i) There exists a universal constant M_0 such that

$$\begin{aligned} &\Big[E_{G_0}\{t^*_G(Y;\rho)-\xi\}^2-R^*(G_0)\Big]^{1/2} \\ \le\ & M_0\max\Big\{|\log\rho|^{3/2},|\log(d(f_G,f_{G_0}))|^{1/2}\Big\}d(f_G,f_{G_0}) \\ &+\Big\{\int\Big(1-\frac{f_{G_0}}{\rho}\Big)_+^2\frac{(f'_{G_0})^2}{f_{G_0}}\Big\}^{1/2}, \end{aligned} \tag{2.27}$$

where $R^*(G_0)=E_{G_0}\{t^*_{G_0}(Y)-\xi\}^2$ is the minimum Bayes risk in (2.7).

(ii) If $\int_{|u|>x_0}G_0(du)\le M_1|\log\rho|^3\epsilon_0^2$ and $2(x_0+1)\rho\le M_2|\log\rho|^2\epsilon_0^2$ for a certain $\epsilon_0\ge d(f_G,f_{G_0})$ and finite positive constants $\{x_0,M_1,M_2\}$, then

$$\begin{aligned} &E_{G_0}\{t^*_G(Y;\rho)-\xi\}^2-R^*(G_0) \\ \le\ & 2(M_0+M_1+M_2)\max\Big(|\log\rho|^3,|\log\epsilon_0|\Big)\epsilon_0^2, \end{aligned} \tag{2.28}$$

where M_0 is universal constant.

Remark 2.2. *For $G=G_0$ (2.27) becomes an identity, so that the square of the first term on the right-hand side of (2.27) represents an upper bound for the regret of using a misspecified G in the regularized Bayes estimator (2.23) instead of the true G_0 for the same regularization level ρ. Under the additional tail probability condition on G_0 and for sufficiently small ρ, (2.28) provides an upper bound for the regret of not knowing G_0, compared with the Bayes estimator (2.22) with the true $G=G_0$.*

Remark 2.3. *Since the second term on the right-hand side of (2.27) is increasing in ρ and the first is logarithmic in $1/\rho$, we are allowed to take $\rho>0$ of much smaller order than $d(f_G,f_{G_0})$ in (2.27), for example, under moment conditions on G_0. Still, the cubic power of the logarithmic factors in (2.27) and (2.28) is crude.*

The following lemma plays a crucial role in the proof of Theorem 2.1.

Lemma 2.1. *Let $d(f,g)$ be as in (2.26) and $\widetilde{L}(y)=\sqrt{-\log(2\pi y^2)}$. Then,*

$$\int \frac{(f'_G - f'_{G_0})^2}{f_G \vee \rho + f_{G_0} \vee \rho} \le 2e^2 d^2(f_G, f_{G_0}) \max(\widetilde{L}^6(\rho), 2a^2) \tag{2.29}$$

for $\rho \le 1/\sqrt{2\pi}$, where $a^2 = \max\{\widetilde{L}^2(\rho)+1, |\log d^2(f_G, f_{G_0})|\}$.

Proof of Theorem 2.1. Let

$$\|g\|_h = \Big\{ \int g^2(x)h(x)dx \Big\}^{1/2}$$

be the $L_2(h(x)dx)$ norm for $h \ge 0$. Since $t^*_{G_0}$ is the Bayes rule, by (2.23)

$$\begin{aligned} & [E_{G_0}\{t^*_G(Y;\rho)-\xi\}^2 - E_{G_0}\{t^*_{G_0}(Y)-\xi\}^2]^{1/2} \\ = \;& \|f'_G/(f_G \vee \rho) - f'_{G_0}/f_{G_0}\|_{f_{G_0}} \\ \le \;& r(f_G,\rho) + \|(1-f_{G_0}/\rho)_+ f'_{G_0}/f_{G_0}\|_{f_{G_0}}, \end{aligned} \tag{2.30}$$

where $r(f_G,\rho) = \|f'_G/(f_G \vee \rho) - f'_{G_0}/(f_{G_0} \vee \rho)\|_{f_{G_0}}$.

Let $w_* = 1/(f_G \vee \rho + f_{G_0} \vee \rho)$. For $G_1 = G$ or $G_1 = G_0$,

$$\begin{aligned} \int \Big(\frac{f'_{G_1}}{f_{G_1} \vee \rho} - 2f'_{G_1} w_* \Big)^2 f_{G_0} &\le \int \Big(\frac{f'_{G_1}}{f_{G_1} \vee \rho} |f_G - f_{G_0} w_* \Big)^2 f_{G_0} \\ &\le \widetilde{L}^2(\rho) \int (f_G - f_{G_0})^2 w_*^2 f_{G_0} \end{aligned}$$

due to $|f'_{G_1}|/(f_{G_1} \vee \rho) \le \widetilde{L}(\rho)$ by (2.24). Since $(\sqrt{f_G} + \sqrt{f_{G_0}})^2 w_* \le 2$ and $w_* f_{G_0} \le 1$, we find

$$\begin{aligned} r(f_G,\rho) &\le 2\|(f'_G - f'_{G_0})w_*\|_{f_{G_0}} + 2\widetilde{L}(\rho)\|(f_G - f_{G_0})w_*\|_{f_{G_0}} \\ &\le 2\|(f'_G - f'_{G_0})\|_{w_*} + 2\widetilde{L}(\rho)\sqrt{2} d(f_G, f_{G_0}). \end{aligned}$$

Thus, (2.27) follows from (2.29) and (2.30).

To prove (2.28) we use Lemma 6.1 in [76]:

$$\begin{aligned} & \int_{f_{G_0}} \Big(\frac{f'_{G_0}}{f_{G_0}} \Big)^2 f_{G_0} \\ \le \;& \int_{|u|>x_0} G_0(du) + 2x_0\rho \max\{\widetilde{L}^2(\rho), 2\} + 2\rho\sqrt{\widetilde{L}^2(\rho)+2} \\ \le \;& (M_1 + M_2)|\log \rho|^3 \epsilon_0^2, \end{aligned}$$

due to $|\log\rho| \ge \widetilde{L}^2(\rho) \ge 2$. This and (2.27) imply (2.28). □

2.4.2 An oracle inequality for the GMLEB

In this subsection, we provide an oracle inequality which bounds the regret (2.11) of using the GMLEB $t^*_{\widehat{G}_n}$ in (2.15) against the oracle Bayes rule $t^*_{G_n}$ in (3.16).We provide the main elements leading to the oracle inequality before presenting the oracle inequality and an outline of its proof.

It follows from the fundamental theorem of compound decisions (2.5) that for separable estimators $\widehat{\boldsymbol{\theta}} = t(\boldsymbol{X})$, the compound risk is identical to the MSE of $\widehat{\xi} = t(Y)$ for the estimation of a single real random parameter xi under P_G in (2.3), so that Theorem 2.1 provides an upper bound for the regret of the regularized Bayes rule $t^*_G(\boldsymbol{X};\rho)$ in terms of the Hellinger distance $d(f_G, f_{G_n})$ and $\rho > 0$. There is a large deviation upper bound for the Hellinger distance $d(f_{\widehat{G}_n}, f_{G_n})$ in [80]. We will show that the GMLEB estimator $t^*_{\widehat{G}_n}(\boldsymbol{X})$ is identical to its regularized version $t^*_{\widehat{G}_n}(\boldsymbol{X};\rho_n)$ for certain $|\log \rho_n| \asymp \log n$ when the generalized MLE (2.12) or its approximation (2.14) are used. Still, $t^*_{\widehat{G}_n}(\boldsymbol{X};\rho_n)$ is not separable, since the generalized MLE $\widehat{G}_n$ is based on the same data $\boldsymbol{X}$. A natural approach of deriving oracle inequalities is then to combine Theorem 2.1 with a maximal inequality. This requires in addition an entropy bound for the class of regularized Bayes rules $t^*_G(x;\rho)$ with given $\rho > 0$ and an exponential inequality for the difference between the loss and risk for each regularized Bayes rule. In the rest of this subsection, we provide these crucial components of our theoretical investigation.

(a) *A large deviation inequality for the convergence of an approximate generalized MLE.* Under the iid assumption of the EB model (2.48), Ghosal and van der Vaart [39] obtained an exponential inequality for the Hellinger loss of the generalized MLE of a normal mixture density in terms of the L_∞ norm of θ_i. This result can be improved upon using their newer entropy calculation in [40]. The results in [39, 40] are unified and further improved upon in the iid case and extended to deterministic $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ in weak ℓ_p balls for all $0 < p \leq \infty$ in [80]. This latest result, stated below as Theorem 2.2, will be used here in conjunction of Theorem 2.1 to prove oracle inequalities for the GMLEB.

The p-th weak moment of a distribution G is

$$\mu_p^w(G) = \Big\{ \sup_{x>0} x^p \int_{|u|>x} G(du) \Big\}^{1/p} \tag{2.31}$$

with $\mu_\infty^w(G) = \inf\{x\colon \int_{|u|>x} G(du) = 0\}$. Define convergence rates

$$\begin{aligned} \epsilon(n,G,p) &= \max\Big[\sqrt{2\log n}, \{n^{1/p}\sqrt{\log n}\mu_p^w(G)\}^{p/(2+2p)}\Big]\sqrt{\frac{\log n}{n}} \\ &= \max\Big[\sqrt{\frac{2\log n}{n}}, \Big\{\sqrt{\log n}\frac{\mu_p^w(G)}{n}\Big\}^{p/(2+2p)}\Big]\sqrt{\log n} \end{aligned} \tag{2.32}$$

with $\epsilon(n,G,\infty) = \{(2\log n) \vee (\sqrt{\log n}\mu_\infty^w(G))\}^{1/2}\sqrt{(\log n)/n}$.

Theorem 2.2. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in \mathbb{R}^n$. Let f_G and G_n be as in (2.13) and (2.4), respectively. Let $\widehat{G}_n$ be certain approximate generalized MLE satisfying (2.14). Then, there exists a universal constant x_* such that for all $x \ge x_*$ and $\log n \ge 2/p$,*

$$P_{n,\boldsymbol{\theta}}\Big\{d(f_{\widehat{G}_n}, f_{G_n}) \ge x\epsilon_n\Big\} \le \exp\Big(-\frac{x^2 n\epsilon_n^2}{2\log n}\Big) \le e^{-x^2\log n}, \tag{2.33}$$

where $\epsilon_n = \epsilon(n, G_n, p)$ is as in (2.32) and $d(f,g)$ is the Hellinger distance (2.26). In particular, for any sequences of constants $M_n \to \infty$ and fixed positive α and c,

$$\epsilon_n \asymp \begin{cases} n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}, & \text{if } \mu_p^w(G_n) = O(1) \text{ with a fixed } p, \\ n^{-1/2}(\log n)^{3/4}\{M_n^{1/2} \vee (\log n)^{1/4}\}, & \text{if } G_n([-M_n, M_n]) = 1 \text{ and } p = \infty, \\ n^{-1/2}(\log n)^{1/(2(2\wedge\alpha))+3/4}, & \text{if } \int e^{|cu|^\alpha} G_n(du) = O(1) \text{ and } p \asymp \log n. \end{cases}$$

Remark 2.4. *Under the condition $G([-M_n, M_n]) = 1$ and the iid assumption (2.48) with G depending on n, the large deviation bound in [39] provides the convergence rate $\epsilon_n \asymp n^{-1/2}(\log n)^{1/2}\{M_n \vee (\log n)^{1/2}\}$, and the entropy calculation in [40] leads to the convergence rate $\epsilon_n \asymp n^{-1/2}(\log n)\sqrt{M_n}$. These rates are slower than the rate in Theorem 2.2 when $M_n/\sqrt{\log n} \to \infty$.*

Remark 2.5. *The proof of Theorem 2.2 is identical for the generalized MLE (2.12) and its approximation (2.14). The constant x_* is universal for $q_n = (e\sqrt{2\pi}/n^2) \wedge 1$ in (2.14) and depends on $\sup_n |\log q_n|/\log n$ in general.*

(b) *Representation of the GMLEB estimator as a regularized one at data points.* The connection between the GMLEB estimator (2.15) and the regularized Bayes rule (2.23) in Theorem 2.1 is provided by

$$t^*_{\widehat{G}_n}(\boldsymbol{X}) = t^*_{\widehat{G}_n}(\boldsymbol{X};\rho_n), \quad \rho_n = q_n/(en\sqrt{2\pi}), \tag{2.34}$$

where q_n is as in (2.14). This is consequence of the following proposition.

Proposition 2.2. *Let $f(x|u)$ be a given family of densities and $\{X_i, i \le n\}$ be given data. Let $\widehat{G}_n$ be an approximate generalized MLE of a mixing distribution satisfying*

$$\prod_{i=1}^n \int f(X_i|u)\widehat{G}_n(du) \ge q_n \sup_G \prod_{i=1}^n \int f(X_i|u)G(du)$$

for certain $0 < q_n \le 1$. Then, for all $j = 1, \ldots, n$

$$f_{\widehat{G}_n}(X_j) = \int f(X_j|u)\widehat{G}_n(du) \ge \frac{q_n}{en} \sup_u f(X_j|u).$$

In particular, (2.34) holds for $f(x|u) = \varphi(x-u)$.

Proof of Proposition 2.2. Let j be fixed and $u_j = \arg\max_u f(X_j|u)$. Define $\widehat{G}_{n,j} = (1-\epsilon)\widehat{G}_n + \epsilon\delta_{u_j}$ with $\epsilon = 1/n$, where δ_u is the unit mass at u. Since $f(x|u) \ge 0$, $f_{\widehat{G}_{n,j}}(X_i) \ge (1-\epsilon)f_{\widehat{G}_n}(X_i)$ and $f_{\widehat{G}_{n,j}}(X_j) \ge \epsilon f(X_j|u_j)$, so that

$$\frac{1}{q_n}\prod_{i=1}^n f_{\widehat{G}_n}(X_i) \ge \prod_{i=1}^n f_{\widehat{G}_{n,j}}(X_i) \ge (1-\epsilon)^{n-1}\epsilon f(X_j|u_j)\prod_{i\ne j} f_{\widehat{G}_n}(X_i).$$

Thus, $f_{\widehat{G}_n}(X_j) \ge q_n(1-\epsilon)^{n-1}\epsilon f(X_j|u_j)$ with $\epsilon = 1/n$, after the cancelation of $f_{whG}(X_i)$ for $i \ne j$. The conclusion follows from $(1-1/n)^{n-1} \ge 1/e$. □

(c) *An entropy bound for regularized Bayes rules.* We now provide an entropy bound for collection of regularized Bayes rule. For any family $\mathscr{H}$ of functions and semidistance d_0, the ϵ-covering number is

$$N(\epsilon, \mathscr{H}, d_0) = \inf\Big\{N\colon \mathscr{H} \subseteq \cup_{j=1}^N \mathrm{Ball}(h_j, \epsilon, d_0)\Big\} \tag{2.35}$$

with $\mathrm{Ball}(h, \epsilon, d_0) = \{f\colon d_0(f,h) < \epsilon\}$. For each fixed $\rho > 0$ define the complete collection of the regularized Bayes rules $t^*_G(x;\rho)$ in (2.23) as

$$\mathscr{T}_\rho = \Big\{t^*_G(\cdot;\rho)\colon G \in \mathscr{G}\Big\}, \tag{2.36}$$

where $\mathscr{G}$ is the family of all distribution functions. The following proposition provides an entropy bound for (2.36) under the seminorm $\|h\|_{\infty,M} = \sup_{|x|\le M}|h(x)|$.

Proposition 2.3. *Let* $\widetilde{L}(y) = \sqrt{-\log(2\pi y^2)}$ *be the inverse of* $y = \varphi(x)$ *as in Proposition 2.1. Then, for all* $0 < \eta \le \rho \le (2\pi e)^{-1/2}$,

$$\begin{aligned}&\log N(\eta^*, \mathscr{T}_\rho, \|\cdot\|_{\infty,M})\\ \le\ & \Big\{4\big(6\widetilde{L}^2(\eta)+1\big)\big(2M/\widetilde{L}(\eta)+3\big)+2\Big\}|\log\eta|,\end{aligned} \tag{2.37}$$

where $\eta^* = (\eta/\rho)\{3\widetilde{L}(\eta)+2\}$.

(d) *An exponential inequality for the loss of regularized Bayes rules.* The last element of our proof is an exponential inequality for the difference between the loss and risk of regularized Bayes rules $t^*_G(\boldsymbol{X};\rho)$. For each separable rule $t(x)$, the squared loss $\|t(\boldsymbol{X})-\boldsymbol{\theta}\|^2$ is a sum of independent variables. However, a direct application of the empirical process theory to the loss would yield an oracle inequality of the $n^{-1/2}$ order, which is inadequate for the sharper convergence rates in this chapter. Thus, we use the following isoperimetric inequality for the square root of the loss.

Proposition 2.4. *Suppose* $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ *under* $P_{n,\boldsymbol{\theta}}$. *Let* $t_G(x;\rho)$ *be the regularized Bayes rule as in (2.23), with a deterministic distribution* G *and* $0 < \rho \le (2\pi e^3)^{-1/2}$. *Let* $\widetilde{L}(\rho) = \sqrt{-\log(2\pi\rho^2)}$. *Then, for all* $x > 0$,

$$P_{n,\boldsymbol{\theta}}\Big\{\|t^*_G(\boldsymbol{X};\rho)-\boldsymbol{\theta}\| \ge E_{n,\boldsymbol{\theta}}\|t^*_G(\boldsymbol{X};\rho)-\boldsymbol{\theta}\|+x\Big\} \le \exp\Big(-\frac{x^2}{2\widetilde{L}^4(\rho)}\Big).$$

Proof of Proposition 2.4. Let $h(\boldsymbol{x}) = \|t^*_G(\boldsymbol{x};\rho)-\boldsymbol{\theta}\|$. It follows from Proposition 2.1 that

$$\begin{aligned}|h(\boldsymbol{x})-h(\boldsymbol{y})| &\le \big\|t^*_G(\boldsymbol{x};\rho)-t^*_G(\boldsymbol{y};\rho)\big\|\\ &\le \|\boldsymbol{x}-\boldsymbol{y}\|\sup_x|(\partial/\partial x)t^*_G(x;\rho)| \le \widetilde{L}^2(\rho)\|\boldsymbol{x}-\boldsymbol{y}\|.\end{aligned}$$

Thus, $h(\boldsymbol{x})/\widetilde{L}^2(\rho)$ has the unit Lipschitz norm. The conclusion follows from the Gaussian isoperimetric inequality [6]. See Page 439 of [70]. □

Our oracle inequality for the GMLEB, stated in Theorem 2.3 below, is a key result from a mathematical point of view. It builds upon Theorem 2.1 and 2.2 and Proposition 2.2, 2.3 and 2.4 (the regularized Bayes rules with misspecified prior, generalized MLE of normal mixtures, representation of the GMLEB, entropy bounds and Gaussian concentration inequality) and leads to the adaptive ratio optimality and minimaxity.

Theorem 2.3. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in \mathbb{R}^n$ as in (2.1). Let $L_n(\cdot,\cdot)$ be the average squared loss in (2.2) and $0 < p \leq \infty$. Let $t^*_{\widehat{G}_n}(\boldsymbol{X})$ be the GMLEB estimator (2.15) with an approximate generalized MLE $\widehat{G}_n$ satisfying (2.14). Then, there exists a universal constant M_0 such that for all $\log n \geq 2/p$,*

$$\begin{aligned} \widetilde{r}_{n,\boldsymbol{\theta}}\big(t^*_{\widehat{G}_n}(\boldsymbol{X})\big) &= \sqrt{E_{n,\boldsymbol{\theta}} L_n\big(t^*_{\widehat{G}_n}(\boldsymbol{X}), \boldsymbol{\theta}\big)} - \sqrt{R^*(G_n)} \\ &\leq M_0 \epsilon_n (\log n)^{3/2}, \end{aligned} \tag{2.38}$$

where $R^*(G_n)$ is the minimum risk of all separable estimators as in (2.8) with $G_n = G_{n,\boldsymbol{\theta}}$ as in (2.4), and $\epsilon = \epsilon(n, G_n, p)$ is as in (2.32). In particular, for any sequences of constants $M_n \to \infty$ and fixed positive α and c,

$$\epsilon_n \asymp \begin{cases} n^{-p/(2+2p)}(\log n)^{(2+3p)/(4+4p)}, & \textit{if } \mu_p^w(G_n) = O(1) \textit{ with a fixed } p, \\ n^{-1/2}(\log n)^{3/4}\{M_n^{1/2} \vee (\log n)^{1/4}\}, & \textit{if } G_n([-M_n, M_n]) = 1 \textit{ and } p = \infty, \\ n^{-1/2}(\log n)^{1/(2(2\wedge\alpha))+3/4}, & \textit{if } \int e^{|cu|^\alpha} G_n(du) = O(1) \textit{ and } p \asymp \log n. \end{cases}$$

Remark 2.6. *In the proof of Theorem 2.3, applications of Theorems 2.1 and 2.2 resulted in the leading term for the upper bound in (2.38), while the contributions of other parts of the proof are of smaller order.*

The consequences of Theorem 2.3 upon the adaptive ratio optimality and minimaxity of the GMLEB are discussed in the next two sections. Here is an outline of its proof. The large deviation inequality in Theorem 2.2 and the representation of the GMLEB in (2.34) imply that

$$\big\|t^*_{\widehat{G}_n}(\boldsymbol{X}) - \boldsymbol{\theta}\big\| \leq \big\|t^*_{\widehat{G}_n}(\boldsymbol{X}; \rho_n) - \boldsymbol{\theta}\big\| I_{A_n} + \zeta_{1n}, \quad \rho_n = \frac{q_n}{e\sqrt{2\pi n}}, \tag{2.39}$$

where $A_n = \{d(f_{\widehat{G}_n}, f_{G_n}) \le x^*\epsilon_n\}$ and $\zeta_{1n} = \|t^*_{\widehat{G}_n}(\boldsymbol{X};\rho_n) - \boldsymbol{\theta}\| I_{A_n^c}$ with $x^* = x_* \vee 1$. By (2.22) and Proposition 2.1, $|t^*_G(X_i;\rho_n) - \theta_i| \le \widetilde{L}(\rho_n) + |N(0,1)|$, so that Theorem 2.2 provides an upper bound for $E_{n,\boldsymbol{\theta}}\zeta^2_{1n}$. By the entropy bound in Proposition 2.3, there exists a finite collection of distributions $\{H_j, j \le N\}$ of manageable size N such that

$$\zeta_{2n} = \Big\{ \|t^*_{\widehat{G}_n}(\boldsymbol{X};\rho_n) - \boldsymbol{\theta}\| I_{A_n} - \max_{j\le N} \|t^*_{H_j}(\boldsymbol{X};\rho_n) - \boldsymbol{\theta}\| \Big\}_+ \tag{2.40}$$

is small and $d(f_{H_j}, f_{G_n}) \le x^*\epsilon_n$ for all $j \le N$. Since the regularized Bayes rules $t^*_{H_j}(\boldsymbol{X};\rho_n)$ are separable and the collection $\{H_j, j \le N\}$ is of manageable size, the large deviation inequality in Proposition 2.4 implies that

$$\zeta_{3n} = \max_{j\le N} \Big\{ \|t^*_{H_j}(\boldsymbol{X};\rho_n) - \boldsymbol{\theta}\| - E_{n,\boldsymbol{\theta}}\|t^*_{H_j}(\boldsymbol{X};\rho_n) - \boldsymbol{\theta}\| \Big\}_+ \tag{2.41}$$

is small. Since $d(f_{H_j}, f_{G_n}) \le x^*\epsilon_n$, Theorem 2.1 implies that

$$\zeta_{4n} = \max_{j\le N} \sqrt{E_{n,\boldsymbol{\theta}}\|t^*_{H_j}(\boldsymbol{X};\rho_n) - \boldsymbol{\theta}\|^2} - \sqrt{nR^*(G_n)} \tag{2.42}$$

is no greater than $O(x^*\epsilon_n)(\log\rho_n)^{3/2}$, where $R^*(G_n)$ is the general EB benchmark risk in (2.8). Finally, upper bounds for individual pieces $E_{n,\boldsymbol{\theta}}\zeta^2_{jn}$ are put together via

$$\sqrt{E_{n,\boldsymbol{\theta}}\big\|t^*_{\widehat{G}_n}(\boldsymbol{X}) - \boldsymbol{\theta}\big\|^2} \le \sqrt{nR^*(G_n)} + \sqrt{E_{n,\boldsymbol{\theta}}\Big(\sum_{j=1}^4 |\zeta_{jn}|\Big)^2}. \tag{2.43}$$

2.4.3 Adaptive ratio optimality

We discuss here the adaptive ratio optimality of the GMLEB as consequences of the oracle inequality in Theorem 2.3.

The adaptive ratio optimality holds for an estimator $\widehat{\boldsymbol{\theta}}\colon \boldsymbol{X} \to \mathbb{R}^n$ if its risk is uniformly within a fraction of the general EB benchmark

$$\sup_{\boldsymbol{\theta}\in\Theta^*_n} \frac{E_{n,\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})}{R^*(G_{n,\boldsymbol{\theta}})} \le 1 + o(1) \tag{2.44}$$

in certain classes $\Theta^*_n \subset \mathbb{R}^n$ of the unknown vector $\boldsymbol{\theta}$, where $L_n(\cdot,\cdot)$ is the average squared loss (2.2), $G_{n,\boldsymbol{\theta}} = G_n$ is the empirical distribution of the unknowns in

(2.4) and $R^*(G_n)$ is the general EB benchmark risk (2.8) achieved by the oracle Bayes rule $t^*_{G_n}(\boldsymbol{X})$.

Theorem 2.4. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in \mathbb{R}^n$. Let $t^*_{\widehat{G}_n}(\cdot)$ be the GMLEB estimator (2.15) with an approximate solution $\widehat{G}_n$ satisfying (2.14). Let $G_n = G_{n,\boldsymbol{\theta}}$ and $R^*(G)$ be as in (2.4) and (2.7). Then,*

$$\frac{E_{n,\boldsymbol{\theta}} L_n(t^*_{\widehat{G}_n}(\boldsymbol{X}), \boldsymbol{\theta})}{R^*(G_n)} = \frac{E_{n,\boldsymbol{\theta}} \|t^*_{\widehat{G}_n}(\boldsymbol{X}) - \boldsymbol{\theta}\|^2}{\min_t E_{n,\boldsymbol{\theta}} \|t(\boldsymbol{X}) - \boldsymbol{\theta}\|^2} \le 1 + o(1) \tag{2.45}$$

for the compound loss (2.2), provided that for certain constants b_n

$$\frac{nR^*(G_n)}{(\sqrt{\log n} \vee \max_{i\le n} |\theta_i - b_n|)(\log n)^{9/2}} \to \infty.$$

In particular, if $\max_{i\le n} |\theta_i - b_n| = O(\sqrt{\log n})$ and $nR^*(G_n)/(\log n)^5 \to \infty$, then (2.45) holds.

For any sequences of constants $M_n \to \infty$, Theorem 2.4 provides the adaptive ratio optimality (2.44) of the GMLEB in the classes

$$\Theta^*_n = \Big\{\boldsymbol{\theta} \in \mathbb{R}^n \colon R^*(G_{n,\boldsymbol{\theta}}) \ge M_n n^{-1} (\log n)^{9/2} (\sqrt{\log n} \vee \|\boldsymbol{\theta}\|_\infty)\Big\}.$$

This is a consequence of an oracle inequality for the GMLEB $\widehat{t}_n = t^*_{\widehat{G}_n}$ in Theorem 2.3, which uniformly bound from the above

$$\widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n) = \sqrt{E_{n,\boldsymbol{\theta}} L_n\big(\widehat{t}_n(\boldsymbol{X}), \boldsymbol{\theta}\big)} - \sqrt{R^*(G_n)} \tag{2.46}$$

in term of the weak ℓ_p norm of $\boldsymbol{\theta}$. The quantity (2.46) can be viewed as the regret for the minimization of the squared root of the MSE, instead of (2.11). Clearly, $r_{n,\boldsymbol{\theta}}(\widehat{t}_n)/R^*(G_n) \le o(1)$ iff $\widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n)/\sqrt{R^*(G_n)} \le o(1)$.

In the EB literature, the asymptotic optimality of $\widehat{\boldsymbol{\theta}}$ is defined as

$$G_n \xrightarrow{\mathrm{D}} G \ \Rightarrow\ E_{n,\boldsymbol{\theta}} L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) - R^*(G_n) \to 0 \tag{2.47}$$

for deterministic vectors $\boldsymbol{\theta} \in \mathbb{R}^n$ [59, 74]. In the EB model

$$(Y_i, \xi_i) \text{ iid},\ Y_i|\xi_i \sim N(\xi, 1),\ \xi_i \sim G,\ \text{under } P_G, \tag{2.48}$$

with data $\{Y_i\}$, the EB asymptotic optimality is defined as

$$\lim_{n\to\infty} E_G \sum_{i=1}^n (\widehat{\xi}_i - \xi_i)^2/n = R^*(G). \tag{2.49}$$

We call (2.44) adaptive ratio optimality since it is much stronger than both notions of asymptotic optimality in its uniformity in $\boldsymbol{\theta} \in \Theta_n^*$ and its focus on the harder standard of the relative error, due to $R^*(G_n) \le E_{n,\boldsymbol{\theta}} L_n(\boldsymbol{X}, \boldsymbol{\theta}) = 1$. The difference among these optimality properties is significant for moderate samples in view of some very small $R^*(G_n) \approx$ Oracle/1000 in Table 2.1.

Theorem 2.4 is location invariant, since the GMLEB is location equivalent by (2.16) and $R^*(G_n)$ is location invariant by (2.8). Thus, if $\theta_i = b_n$ for most $i \le n$, the GMLEB performs equally well whether $b_n = 0$ or not. Moreover, if $\theta_i \in B\ \forall i$ for a finite set $B \subset \mathbb{R}$, the GMLEB adaptively shrinks towards the points in B [38]. This is evident in Table 2.1 for $\#\{i\colon \theta_i = 7\} \in \{50, 500\}$ with $B = \{0, 7\}$. In fact, if $\#\{x\colon x \in B_n\} = O(1)$ and $\min_{B_n \ni x \ne y \in B_n} |x - y| \to \infty$, then $G_n(B_n) = 1$ implies $R^*(G_n) \to 0$. Threshold methods certainly do not possess these location invariance and multiple shrinkage properties.

We state a more general version of Theorem 2.4. Theorem 2.3 immediately implies the adaptive ratio optimality (2.44) of the GMLEB in the classes $\Theta_n^* = \Theta_n^*(M_n)$ for any sequences of constants $M_n \to \infty$, where

$$\Theta_n^*(M) = \Big\{\boldsymbol{\theta} \in \mathbb{R}^n\colon R^*(G_{n,\boldsymbol{\theta}}) \ge M(\log n)^3 \inf_{p \ge 2/\log n} \epsilon^2(n, G_{n,\boldsymbol{\theta}}, p)\Big\} \tag{2.50}$$

with $G_{n,\boldsymbol{\theta}} = G_n$ as in (2.4) and $\epsilon(n, G, p)$ as in (2.32). This formally stated in the theorem below.

Theorem 2.5. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in \mathbb{R}^n$. Let $t^*_{\widehat{G}_n}(\cdot)$ be the GMLEB estimator (2.15) with the approximate MLE $\widehat{G}_n$ in (2.14). Let $R^*(G_{n,\boldsymbol{\theta}})$ be the general EB benchmark in (2.8) with the distribution $G_n = G_{n,\boldsymbol{\theta}}$ in (2.4). Then for the classes $\Theta_n^*(M)$ in (2.50),*

$$\lim_{(n,M)\to(\infty,\infty)} \sup_{\boldsymbol{\theta} \in \Theta_n^*(M)} \Big\{E_{n,\boldsymbol{\theta}} L_n\big(t^*_{\widehat{G}_n}(\boldsymbol{X}), \boldsymbol{\theta}\big)/R^*(G_{n,\boldsymbol{\theta}})\Big\} \le 1. \tag{2.51}$$

Remark 2.7. *Since the minimum of $\epsilon(n, G_{n,\boldsymbol{\theta}}, p)$ is taken in (2.50) over $p \geq 2/\log n$ for each $\boldsymbol{\theta}$, the adaptive ratio optimality (2.51) allows smaller $R^*(G_{n,\boldsymbol{\theta}})$ than simply using $\epsilon(n, G_{n,\boldsymbol{\theta}}, \infty)$ does as in Theorem 2.4. Thus, Theorem 2.5 implies Theorem 2.4.*

2.4.4 Adaptive minimaxity

Another main consequence of the oracle inequality in Theorem 2.3 is the adaptive minimaxity of the GMLEB for a broad range of sequences $\boldsymbol{\theta} \in \mathbb{R}^n$.

Minimaxity is commonly used to measure the performance of statistical procedures. For $\Theta \in \mathbb{R}^n$, the minimax risk for the average squared loss (2.2) is

$$\mathscr{R}_n(\Theta) = \inf_{\widetilde{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta} \in \Theta} E_{n,\boldsymbol{\theta}} L_n(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}), \tag{2.52}$$

where the infimum is taken over all Borel mappings $\widetilde{\boldsymbol{\theta}} \colon \boldsymbol{X} \to \mathbb{R}^n$. An estimator is minimax in a specific class Θ of unknown mean vectors if it attains $\mathscr{R}_n(\Theta)$, but this does not guarantee satisfactory performance since the minimax estimator is typically uniquely tuned to the specific set Θ. For small Θ, the minimax estimator has high risk outside Θ. For large Θ, the minimax estimator is too conservative by focusing on the worst case scenario within Θ. Adaptive minimaxity overcomes this difficulty by requiring

$$\frac{\sup_{\boldsymbol{\theta} \in \Theta_n} E_{n,\boldsymbol{\theta}} L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})}{\mathscr{R}_n(\Theta_n)} \to 1 \tag{2.53}$$

uniformly for a wide range of sequences $\{\Theta_n \subset \mathbb{R}^n, n \geq 1\}$ of parameter classes. Define (regular or strong) ℓ_p balls as

$$\Theta_{p,C,n} = \Big\{ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \colon n^{-1} \sum_{i=1}^{n} |\theta_i|^p \leq C^p \Big\}. \tag{2.54}$$

The quantity C in (2.54), called length-normalized or standardized radius of the ℓ_p ball, is denoted as η in [1, 21, 47], where adaptive minimaxity in ℓ_p balls with $C = C_n \to 0$ and $p < 2$ is used to measure the performance of estimators for sparse $\boldsymbol{\theta}$. The following theorem establishes the adaptive minimaxity of the GMLEB in

ℓ_p balls with radii $C = C_n$ in intervals diverging to $(0, \infty)$. This covers sparse and dense $\boldsymbol{\theta}$ simultaneously.

Theorem 2.6. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in \mathbb{R}^n$. Let $\widehat{\boldsymbol{\theta}} = t^*_{\widehat{G}_n}(\boldsymbol{X})$ be the GMLEB in (2.15) with an approximate solution $\widehat{G}_n$ satisfying (2.14). Let $L_n(\cdot,\cdot)$ be the average squared loss (2.2) and $\mathscr{R}_n(\Theta)$ be the minimax risk (2.52). Then, as $n \to \infty$, the adaptive minimaxity (3.34) holds in ℓ_p balls (2.54) with $\Theta_n = \Theta_{p,C_n,n}$, provided that*

$$\frac{n^{1/(p\wedge 2)}C_n}{(\log n)^{\kappa_1(p)}} \to \infty, \quad \frac{C_n}{n}(\log n)^{\kappa_2(p)} \to 0, \tag{2.55}$$

where $\kappa_1(p) = 1/2 + 4/p + 3/p^2$ for $p < 2$, $\kappa_1(2) = 13/4$, $\kappa_1(p) = 5/2$ for $p > 2$, and $\kappa_2(p) = 9/2 + 4/p$.

Theorem 2.6 is a consequence of the oracle inequality (2.38) and the minimax theory in [21]. An outline of this argument is given in this subsection. An alternative statement of the conclusion of Theorem 2.6 is

$$\lim_{(n,M)\to(\infty,\infty)} \sup_{C\in\mathscr{C}_{p,n}(M)} \frac{\sup_{\boldsymbol{\theta}\in\Theta_{p,C,n}} E_{n,\boldsymbol{\theta}} L_n\big(t^*_{\widehat{G}_n}(\boldsymbol{X}), \boldsymbol{\theta}\big)}{\mathscr{R}_n(\Theta_{p,C,n})} = 1$$

where $\mathscr{C}_{p,n}(M) = [Mn^{-1/(p\wedge 2)}(\log n)^{\kappa_1(p)}, n/\{M(\log n)^{\kappa_2(p)}\}]$. The powers $\kappa_1(p)$ and $\kappa_2(p)$ of the logarithmic factors in (2.55) and in the definition of $\mathscr{C}_{p,n}(M)$ are crude.

Adaptive and approximate minimax estimators of the normal means in ℓ_p balls have been considered in [1, 5, 21, 23, 47, 74, 76]. Donoho and Johnstone [23] proved that as $(n, C_n) \to (\infty, 0+)$, with $nC_n^p/(\log n)^{p/2} \to \infty$ for $p < 2$,

$$\mathscr{R}_n(\Theta_{p,C_n,n}) = (1 + o(1)) \min_{t\in\mathscr{D}} \max_{\boldsymbol{\theta}\in\Theta_{p,C_n,n}} E_{n,\boldsymbol{\theta}} L_n(t(\boldsymbol{X}), \boldsymbol{\theta}), \tag{2.56}$$

where $\mathscr{D}$ is the collection of all (soft and hard) threshold rules. Therefore, adaptive minimaxity (3.34) in small ℓ_p balls $\Theta_n = \Theta_{p,C_n,n}$ can be achieved by threshold rules with suitable data-driven threshold levels. This has been done using the FDR [1] for $(\log n)^5/n \le C_n^p \le n^{-\kappa}$ with $p < 2$ and any $\kappa > 0$. Zhang [76] proved that (3.34) holds for the Fourier general EB estimator of [74] in $\Theta_n = \Theta_{p,C_n,n}$ for $C_n^p\sqrt{n}/(\log n)^{1+(p\wedge 2)/2} \to \infty$.

A number of estimators have been proven to possess the adaptive rate minimaxity in the sense of attaining within a bounded factor of the minimax risk. In ℓ_p balls $\Theta_{p,C_n,n}$, the EBThresh is adaptive rate minimax for $p \leq 2$ and $nC_n^p \geq (\log n)^2$ [47], while the generalized C_p is adaptive rate minimax for $p < 2$ and $1 \leq O(1)nC_n^p$ [5]. It follows from [76] that a hybrid between the Fourier general EB and universal soft threshold estimators is also adaptive rate minimax in $\Theta_{p,C_n,n}$ for $1 \leq O(1)nC_n^p$.

The adaptive minimaxity as provided in Theorem 2.6 unifies the adaptive minimaxity of different types estimators in different ranges of the radii C_n of the ℓ_p balls with the exception of the two very extreme ends, due to the crude power $\kappa_1(p)$ of the logarithmic factor for small C_n and the requirement of an upper bound for large C_n. The hybrid Fourier general EB estimator achieves the adaptive rate minimaxity in a wider range of ℓ_p balls than what we prove here for the GMLEB. However, as we have seen in Table 2.1, the finite sample performance of the GMLEB is much stronger. It seems that the less stringent and commonly considered adaptive rate minimaxity leaves too much room to provide adequate indication of finite sample performance.

Instead of the general EB approach, adaptive minimax estimation in small ℓ_p balls can be achieved by threshold methods, provided that the radius is not too small. However, since (2.56) does not hold for fixed $p > 0$ and $C \in (0, \infty)$, threshold estimators are not asymptotically minimax with $\Theta_n = \Theta_{p,C,n}$ in (3.34) for fixed (p, C). Consequently, adaptive minimax estimations in small, fixed and large ℓ_p balls are often treated separately in the literature. We now explain the general EB approach for adaptive minimax estimation which provides a unified treatment for ℓ_p balls of different ranges of radii. This provides an outline for the proof of Theorem 2.6.

We first discuss the relationship between the minimax estimation of a deterministic vector $\boldsymbol{\theta}$ in ℓ_p balls and the minimax estimation of a single random mean under an unknown "prior" in L_p balls. For positive p and C, the L_p balls of distribution functions are defined as

$$\mathscr{G}_{p,C} = \Big\{G\colon \int |u|^p G(du) \le C^p\Big\}.$$

Since $\mathscr{G}_{p,C}$ is a convex class of distributions, the minimax theorem provides

$$\mathscr{R}(\mathscr{G}_{p,C}) = \min_t \max_{G\in\mathscr{G}_{p,C}} E_G(t(Y)-\xi)^2 = \max_{G\in\mathscr{G}_{p,C}} R^*(G) \le 1 \tag{2.57}$$

for the estimation of a single real random parameter ξ in the model (2.3), where $R^*(G)$ is the minimum Bayes risk in (2.7). Thus, since $G_n = G_{n,\boldsymbol{\theta}} \in \mathscr{G}_{p,C}$ for $\boldsymbol{\theta} \in \Theta_{p,C,n}$, the fundamental theorem of compound decisions (2.5) implies that (2.57) dominates the compound minimax risk (2.52) in ℓ_p balls:

$$\mathscr{R}_n(\Theta_{p,C,n}) \le \inf_{t(x)} \sup_{\boldsymbol{\theta}\in\Theta_{p,C,n}} E_{n,\boldsymbol{\theta}} L_n(t(\boldsymbol{X}),\boldsymbol{\theta}) \le \mathscr{R}(\mathscr{G}_{p,C}) \le 1. \tag{2.58}$$

Donoho and Johnstone [21] proved that as $C^{p\wedge 2} \to 0+$

$$\Big|\frac{\mathscr{R}(\mathscr{G}_{p,C})}{C^{p\wedge 2}\{2\log(1/C^p)\}^{(1-p/2)_+}} - 1\Big| \to 0 \tag{2.59}$$

and that for either $p \ge 2$ with $C_n > 0$ or $p < 2$ with $nC_n^p/(\log n)^{p/2} \to \infty$,

$$\Big|\frac{\mathscr{R}_n(\Theta_{p,C_n,n})}{\mathscr{R}(\mathscr{G}_{p,C_n})} - 1\Big| \to 0. \tag{2.60}$$

In the general EB approach, the aim is to find an estimator $\widehat{t}_n$ of $t^*_{\widehat{G}_n}$ with small regret (2.11) or (2.46). If the approximation to $t^*_{\widehat{G}_n}$ in risk is sufficiently accurate and uniformly within a small fraction of $\mathscr{R}(\mathscr{G}_{p,C_n})$ for $\boldsymbol{\theta} \in \Theta_{p,C_n,n}$, the maximum risk of the general EB estimator in $\Theta_{p,C_n,n}$ would be within the same small fraction of $\mathscr{R}(\mathscr{G}_{p,C_n})$, since the risk of $t^*_{\widehat{G}_n}$ is bounded by $R^*(G_{n,\boldsymbol{\theta}}) \le \mathscr{R}(\mathscr{G}_{p,C_n})$ for $\boldsymbol{\theta} \in \Theta_{p,C_n,n}$. Thus, (2.60) plays a crucial role in general EB.

It follows from (2.46), (2.57) and (2.54) that

$$\sup_{\boldsymbol{\theta}\in\Theta_{p,C,n}} \sqrt{E_{n,\boldsymbol{\theta}} L_n\big(\widehat{t}_n(\boldsymbol{X}),\boldsymbol{\theta}\big)} \le \sup_{\boldsymbol{\theta}\in\Theta_{p,C,n}} \widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n) + \sqrt{\mathscr{R}(\mathscr{G}_{p,C})}. \tag{2.61}$$

Thus, by (2.59) and (2.60), the adaptive minimaxity (3.34) of $\widehat{\boldsymbol{\theta}} = \widehat{t}_n(\boldsymbol{X})$ in ℓ_p balls $\Theta_n = \Theta_{p,C_n,n}$ is a consequence of an oracle inequality of the form

$$\sup_{\boldsymbol{\theta}\in\Theta_{p,C_n,n}} \widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n) = o(1)\sqrt{J_{p,C_n}} \tag{2.62}$$

with $J_{p,C} = \min\{1, C^{p\wedge 2}\{1 \vee (2\log(1/C^p))\}^{(1-p/2)_+}\}$. In our proof, (2.59) and the upper bound $\mathscr{R}(\mathscr{G}_{p,C}) \le 1$ provide $\inf_C \mathscr{R}(\mathscr{G}_{p,C})/J_{p,C} > 0$. Although $J_{p,C}$ provides the order of $\mathscr{R}(\mathscr{G}_{p,C})$ for each p via (2.59), explicit expressions of the minimax risk $\mathscr{R}_n(\Theta_{p,C,n})$ for general fixed (p, C, n) or the minimax risk $\mathscr{R}(\mathscr{G}_{p,C})$ for fixed (p, C) with $p \ne 2$ are still open problems.

We have stated our results for regular ℓ_p balls in Theorem 2.6. In the rest of the subsection, we consider weak ℓ_p balls

$$\Theta^w_{p,C,n} = \Big\{\boldsymbol{\theta} \in \mathbb{R}^n \colon \mu^w_p(G_{n,\boldsymbol{\theta}}) \le C\Big\}, \tag{2.63}$$

where $G_{n,\boldsymbol{\theta}}$ is the empirical distribution of the components of $\boldsymbol{\theta}$ and the function $\mu^w_p(G)$ is the weak moment in (2.31). Alternatively,

$$\Theta^w_{p,C,n} = \Big\{\boldsymbol{\theta} \in \mathbb{R}^n \colon \max_{1\le i\le n} |\theta_i|^p \sum_{j=1}^n I\{|\theta_j| \ge |\theta_i|\}/n \le C^p\Big\}.$$

Theorem 2.7. *Let $\boldsymbol{X} \sim N(\boldsymbol{\theta}, \boldsymbol{I}_n)$ under $P_{n,\boldsymbol{\theta}}$ with a deterministic $\boldsymbol{\theta} \in \mathbb{R}^n$. Let $L_n(\cdot,\cdot)$ be the average squared loss (2.2) and $\mathscr{R}_n(\Theta)$ be the minimax risk (2.52). Then, for all approximate solutions $\widehat{G}_n$ satisfying (2.14), the GMLEB $\widehat{\boldsymbol{\theta}} = t^*_{\widehat{G}_n}(\boldsymbol{X})$ is adaptive minimax (3.34) in the weak ℓ_p balls $\Theta_n = \Theta^w_{p,C_n,n}$ in (2.63), provided that the radii C_n are within the range (2.55).*

Here is our argument. The weak L_p balls that matches (2.63) is

$$\mathscr{G}^w_{p,C} = \Big\{G \colon \mu^w_p(G) \le C\Big\}.$$

Let $J^w_{p,C}(\lambda) = -\int_0^\infty (t^2 \wedge \lambda^2) d\{1 \wedge (C/t)^p\}$, which is approximately the Bayes risk of the soft threshold estimator for the stochastically largest Pareto Prior in $\mathscr{G}_{p,C}$. Let $\lambda_{p,C} = \sqrt{1 \vee \{2\log(1/C^{p\wedge 2})\}}$. Johnstone [46] proved that

$$\lim_{n\to\infty} \frac{\mathscr{R}_n(\Theta^w_{p,C_n,n})}{\mathscr{R}(\mathscr{G}^w_{p,C_n})} = 1 \tag{2.64}$$

for $p > 2$ with $C_n \to C+ \ge 0$ and for $p \le 2$ with $nC^p_n/(\log n)^{1+6/p} \to \infty$, and that $\mathscr{R}(\mathscr{G}^w_{p,C_n})/J^w_{p,C_n}(\lambda_{p,C_n}) \to 1$ as $C^{p\wedge 2}_n \to 0$. Abramovich et al. [1] proved $\mathscr{R}_n(\Theta^w_{p,C_n,n})/J^w_{p,C_n}(\lambda_{p,C_n}) \to 1$ for $p < 2$ and $(\log n)^5/n \le C^p_n \le n^{-\kappa}$ for all $\kappa > 0$.

The combination of their results implies (2.64) for $p \le 2$ and $C_n^p \ge (\log n)^5/n$. Therefore, (2.64) holds under (2.55) due to $p\kappa_1(p) = p/2 + 4 + 3/p > 5$ for $p < 2$. As in Section (2.60), the adaptive minimaxity in weak ℓ_p balls $\Theta^w_{p,C_n,n}$ is a consequence of

$$\sup_{\boldsymbol{\theta}\in\Theta^w_{p,C_n,n}} \widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n) = o(1)\sqrt{J_{p,C_n}} \tag{2.65}$$

as in (2.62), due to $J_{p,C_n} \asymp \mathscr{R}(\mathscr{G}_{p,C_n}) \le \mathscr{R}(\mathscr{G}^w_{p,C_n})$.

2.5 The Fourier General Empirical Bayes Method

2.5.1 The RF-GEB method

Zhang [74] proposed a general EB method based on a Fourier infinite-order smoothing kernel. It is directly derived from (2.23) using the kernel method

$$\widehat{\boldsymbol{\theta}} = \widehat{t}_n(\boldsymbol{X}),\ \widehat{t}_n(x) = x + \frac{\widehat{f}'_n(x)}{\widehat{f}_n(x) \vee \rho_n},\ \widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^n K(X_i - x, a_n) \tag{2.66}$$

where $\widehat{f}_n$ is a kernel estimator of f_{G_n} in (2.13) based on $X_1, \ldots, X_n$ using the Fourier kernel

$$K(x,a) = \frac{1}{2\pi}\int_{-a}^{a} e^{ixt}dt = \begin{cases} \sin(ax)/(\pi x) & \text{if } x \ne 0, \\ a/\pi & \text{if } x = 0. \end{cases} \tag{2.67}$$

We call the estimator (2.66) F-GEB since we use a special Fourier kernel. The estimator (2.66) approximates the oracle regularized Bayes estimator of the form

$$\widehat{\boldsymbol{\theta}} = t^*_{G_n}(\boldsymbol{X};\rho_n), \quad t^*_{G_n}(x;\rho_n) = x + \frac{f'_{G_n}(x)}{f_{G_n}(x) \vee \rho_n} \tag{2.68}$$

where $f_G(x)$ is the normal location mixture density by distribution G as in (2.13). The oracle regularized Bayes estimator (2.68) has the risk

$$E_{n,\boldsymbol{\theta}} L_n(t^*_{G_n}(\boldsymbol{X};\rho_n), \boldsymbol{\theta}) = 1 - J(\rho_n, G_n),$$

where

$$J(\rho, G) = \int_{-\infty}^{\infty} \left\{\frac{f'_G(x)}{f_G(x)}\right\}^2 \left\{2 - \frac{f_G(x)}{f_G(x)\vee\rho}\right\}\left\{\frac{f_G(x)}{f_G(x)\vee\rho}\right\} f_G(x)dx. \tag{2.69}$$

A reason for using the Fourier kernel (2.67) is the extreme thin tail of $f^*_{G_n}(t) = \int e^{ixt} f_{G_n}(x)dx$, bounded by $e^{-t^2/2}$ in absolute value, and

$$\begin{aligned} & E\widehat{f}^{(k)}_n(x) - f^{(k)}_{G_n}(x) \\ = \; & \frac{1}{2\pi}\int_{|t|\le a_n} (-it)^k e^{-ixt} E_{n,\boldsymbol{\theta}} \sum_{i=1}^n \frac{e^{iX_i t}}{n} dt - \frac{1}{2\pi}\int (-it)^k e^{-ixt} f^*_{G_n}(t)dt \\ = \; & -\frac{1}{2\pi}\int_{|t|>a_n} (-it)^k e^{-ixt} f^*_{G_n}(t)dt, \end{aligned}$$

where $h^{(k)} = (\partial/\partial x)^k h$ for any function h if the derivative exists. Zhang [74] proved that the F-GEB (2.66) approximates the risk $1 - J(\rho_n, G_n)$ at the rate of $(\log n)^{3/2}/(\rho_n n)$ uniformly in $\boldsymbol{\theta}$.

Definition 2.1. *Let X be the finite set $\{x_1, \ldots, x_n\}$ with simple order $x_1 < x_2 < \ldots < x_n$. A real valued function h on X is isotonic, if $x_i, x_j \in X$ and $x_i < x_j$ imply $h(x_i) \le h(x_j)$. Let g be a given function on X. An isotonic function $\widetilde{g}$ is an isotonic regression of g with respect to $x_1 < x_2 < \ldots < x_n$ if it minimizes the sum*

$$\sum_{x \in X} \big(g(x) - h(x)\big)^2$$

in the class of all isotonic functions h on X.

Since the oracle Bayes estimator $t^*_{G_n}$ in (3.16) is monotone increasing, we consider the regularized Fourier general empirical Bayes (RF-GEB) estimator

$$\widehat{\boldsymbol{\theta}} = \widetilde{t}_n(\boldsymbol{X}), \quad \widetilde{t}_n(x) \equiv \arg\min_{t\uparrow} \sum_{i=1}^n \big(t(X_i) - \widehat{t}_n(X_i)\big)^2, \tag{2.70}$$

where $\widehat{t}_n(\cdot)$ is the F-GEB as in (2.66). The RF-GEB is the isotonic regression function of the F-GEB with respect to $X_1, X_2, \ldots, X_n$.

2.5.2 An oracle inequality for the RF-GEB

Theorem 2.8 below asserts that the RF-GEB (2.70) approximates the truncated Bayes estimator in risk at the rate of $(\log n)^{3/2}/(\rho_n n)$. In the numerical experiments in Section 2.6, we can see that the RF-GEB has smaller risk than the F-GEB while we can only prove they have same convergence rate here.

Theorem 2.8. *Let $\widehat{\theta}_i = \widetilde{t}_n(X_i)$ be the RF-GEB estimator given by (2.70). Choose $a_n > 0$ and $\rho_n > 0$ such that $\sqrt{2\log n} \le a_n = O(\sqrt{\log n})$ and $a_n/(\rho_n\sqrt{n}) = o(1)$ as $n \to \infty$. Then*

$$E_{n,\boldsymbol{\theta}} L_n(\widetilde{t}_n(\boldsymbol{X}), \boldsymbol{\theta}) \le 1 - J(\rho_n, G_n) + O(1)\frac{(\log n)^{3/2}}{\rho_n n}. \tag{2.71}$$

If we denote the original sequence as $a_1, a_2, \ldots, a_n$, the following "pairwise average" procedure will result in the isotonic regression of $\{a_n\}$ in limit [3].

1. Set $j = 1$. Let $a_{j,i} \leftarrow a_i$, $i = 1, \ldots, n$.

2. Set $i = 1$ and $a_{j+1,1} \leftarrow a_{j,1}$.

3. If $a_{j+1,i} > a_{j,i+1}$, update $a_{j+1,i} = a_{j+1,i+1} \leftarrow (a_{j+1,i} + a_{j,i+1})/2$. If $a_{j+1,i} \le a_{j,i+1}$, update $a_{j+1,i} \leftarrow a_{j+1,i}$ and $a_{j+1,i+1} \leftarrow a_{j,i+1}$.

4. Update $i \leftarrow i+1$, repeat step 3 and stop when $i = n$. We get a new sequence $a_{j+1,1}, a_{j+1,2}, \ldots, a_{j+1,n}$.

5. Update $j \leftarrow j + 1$, repeat step 2-4 again and again.

6. Denote the limit of the sequence $a_{1i}, \ldots, a_{j,i}, \ldots$ as $\widetilde{a}_i$, $i = 1, \ldots, n$. Then the limit sequence $\widetilde{a}_1, \widetilde{a}_2, \ldots, \widetilde{a}_n$ is the isotonic regression of $a_1, a_2, \ldots, a_n$.

Some tedious mathematical exercise shows the following two lemmas:

Lemma 2.2. *Let $a_1 \le a_2$, $b_1 > b_2$, then*

$$\max\left\{\left|a_1 - \frac{b_1 + b_2}{2}\right|, \left|a_2 - \frac{b_1 + b_2}{2}\right|\right\} \le \max\{|a_1 - b_1|, |a_2 - b_2|\}.$$

Lemma 2.3. *Let $a_1 > a_2$, $b_1 > b_2$, then*

$$\left|\frac{a_1 + a_2}{2} - \frac{b_1 + b_2}{2}\right| \le \max\{|a_1 - b_1|, |a_2 - b_2|\}.$$

Define the ℓ_∞ distance between two finite sequences $\{a_i\}$ and $\{b_i\}$, $1 \le i \le n$ as

$$d(\{a_i\}, \{b_i\}) = \max\{|a_i - b_i|, 1 \le i \le n\}.$$

Lemma 2.2 and 2.3 imply that, after each "pairwise average" step, the ℓ_∞ distance between two finite sequences will not increase. Since the isotonic regression is the limit of "pairwise average" procedure, after the isotonic regression, the ℓ_∞ distance between two sequences will not increase. This is formally stated in the next lemma.

Lemma 2.4. *If $\{\widetilde{a}_n\}$ and $\{\widetilde{b}_n\}$ are isotonic regressions of $\{a_n\}$ and $\{b_n\}$, respectively, then*

$$d(\{\widetilde{a}_n\}, \{\widetilde{b}_n\}) \le d(\{a_n\}, \{b_n\}).$$

Proof of Theorem 2.8. We expand the risk of $\widetilde{t}_n$ as follows,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \theta_i\big)^2 \\
= \quad &\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - t^*_{G_n}(X_i)\big)^2 + \frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(t^*_{G_n}(X_i) - \theta_i\big)^2 \\
&+\frac{2}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_n(X_i) - t^*_{G_n}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big) \\
&+\frac{2}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \widehat{t}_n(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big). \qquad (2.72)
\end{aligned}$$

Let g be a real valued function and h be an increasing function, by property of the isotonic regression,

$$\sum_{i=1}^{n}\big(\widetilde{g}(X_i) - h(X_i)\big)^2 \le \sum_{i=1}^{n}\big(g(X_i) - h(X_i)\big)^2. \qquad (2.73)$$

where $\widetilde{g}$ is the isotonic regression of g with respect to $\boldsymbol{X}$. Since $t^*_{G_n}(x) = x + f'_{G_n}(x)/f_{G_n}(x)$ is a non-decreasing function, this and (2.73) imply

$$\sum_{i=1}^{n}\big(\widetilde{t}_n(X_i) - t^*_{G_n}(X_i)\big)^2 \le \sum_{i=1}^{n}\big(\widehat{t}_n(X_i) - t^*_{G_n}(X_i)\big)^2. \qquad (2.74)$$

where $\widetilde{t}_n$ and $\widehat{t}_n$ are the RF-GEB and F-GEB estimators as in (2.70). Thus, by (2.72) and (2.74),

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \theta_i\big)^2 \le{}& \frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_n(X_i) - \theta_i\big)^2 \\
&+\frac{2}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \widehat{t}_n(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big). \qquad (2.75)
\end{aligned}$$

Zhang [74] proved that when $\sqrt{2\log n} \le a_n = O(\sqrt{\log n})$ and $a_n/(\rho_n\sqrt{n}) = o(1)$ as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_n(X_i) - \theta_i\big)^2 \le 1 - J(\rho_n, G_n) + O(1)\frac{(\log n)^{3/2}}{\rho_n n}. \tag{2.76}$$

In order to bound the cross term on the right hand side of (2.75), we need to use decoupleing technique in [74]. Let (Y_n, λ_n) be a random vector independent of (X_i, θ_i), $1 \le i \le n$, such that

$$Y_n|\lambda_n \sim N(\lambda_n, 1), \quad \lambda_n \sim G_n.$$

Let $X_1', \ldots, X_n'$ be random variables such that condition on $\theta_1, \ldots, \theta_n, \lambda_n$, they are independent of $X_1, \ldots, X_n, Y_n$ and distributed according to $X_i' \sim N(\theta_i, 1)$. Define for $1 \le i \le n$,

$$\widehat{t}_{n,[i]}(x) = x + \widehat{f}'_{n,[i]}(x)/\max(\widehat{f}_{n,[i]}(x), \rho_n), \tag{2.77}$$

where $\widehat{f}_{n,[i]}$ is the estimate of f_{G_n} in (2.13) based on $X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_n$,

$$\widehat{f}_{n,[i]}(x) = \frac{1}{n}\Big\{K(X_i' - x, a_n) + \sum_{1\le l\le n, l\ne i} K(X_l - x, a_n)\Big\}. \tag{2.78}$$

Let $\widetilde{t}_{n,[i]}(x)$ be the isotonic regression of $\widehat{t}_{n,[i]}(x)$ with respect to $X_1, \ldots, X_n$. Now we can bound the cross term in (2.75) as follows

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n} E\big(\widetilde{t}_n(X_i) - \widehat{t}_n(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big) \\
= &\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \widetilde{t}_{n,[i]}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big) \\
&+\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_{n,[i]}(X_i) - \widehat{t}_{n,[i]}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big) \\
&+\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_{n,[i]}(X_i) - \widehat{t}_n(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big).
\end{aligned} \tag{2.79}$$

The second term on the right hand side of (2.79) actually vanishes.

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_{n,[i]}(X_i) - \widehat{t}_{n,[i]}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big) \\
= \ & \frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_{n}(X'_i) - \widehat{t}_{n}(X'_i)\big)\big(t^*_{G_n}(X'_i) - \theta_i\big) \\
= \ & E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_{n}(Y_n) - \widehat{t}_{n}(Y_n)\big)\big(t^*_{G_n}(Y_n) - \lambda_n\big) \\
= \ & 0. \qquad (2.80)
\end{aligned}
$$

By Schwarz inequality and the fact that $E_{n,\boldsymbol{\theta}}\big(t^*_{G_n}(X_i) - \theta_i\big)^2 < 1$, for the third term on the right hand side of (2.79),

$$
\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_{n,[i]}(X_i) - \widehat{t}_{n}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big)\right|^2 \\
\leq \ & \frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_{n,[i]}(X_i) - \widehat{t}_{n}(X_i)\big)^2 E_{n,\boldsymbol{\theta}}\big(t^*_{G_n}(X_i) - \theta_i\big)^2 \\
\leq \ & \frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_{n}(X'_i) - \widehat{t}_{n,[i]}(X'_i)\big)^2.
\end{aligned}
$$

Lemma 3 in [74] states that under the conditions of Theorem 2.8,

$$
\left\{\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_{n}(X'_i) - \widehat{t}_{n,[i]}(X'_i)\big)^2\right\}^{1/2} \leq O(1)\frac{a_n^{3/2}}{\rho_n n}.
$$

Hence we have

$$
\left|\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widehat{t}_{n,[i]}(X_i) - \widehat{t}_{n}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big)\right| \leq O(1)\frac{a_n^{3/2}}{\rho_n n}. \qquad (2.81)
$$

We need to work a little bit harder on the first term on the right hand side of (2.79) as it involves the discrepancy between two isotonic sequences. From (2.67), $K'(x) = (ax\cos(ax) - \sin(ax))/(\pi x^2)$. If $x \geq a^{-1}$, $|K'(x)| \leq (2ax)/(\pi x^2) \leq 2a^2/\pi$. Using Taylor expansion, it is easy to see $|K'(x)| = (1+o(1))a^3x/(3\pi)$, hence if $x < a^{-1}$, $|K'(x)| \leq (1+o(1))a^2/(3\pi)$. So we see that $K'(x) \leq O(1)a^2$. Thus we have

$$
\left|\widehat{f}'_n(x) - \widehat{f}'_{n,[i]}(x)\right| = \frac{1}{n}\big|K'(X'_i - x) - K'(X_i - x)\big| \leq O(1)\frac{a_n^2}{n}. \qquad (2.82)
$$

Since $\widehat{f}_n(x) \vee \rho_n \geq \rho_n$ and $\widehat{f}_{n,[i]}(x) \vee \rho_n \geq \rho_n$, by (2.82),

$$\left|\widehat{t}_n(x) - \widehat{t}_{n,[i]}(x)\right| = \left|\frac{\widehat{f}'_n(x)}{\widehat{f}_n(x) \vee \rho_n} - \frac{\widehat{f}'_{n,[i]}(x)}{\widehat{f}_{n,[i]}(x) \vee \rho_n}\right| \leq O(1)\frac{a_n^2}{\rho_n n}. \tag{2.83}$$

By (2.83) and Lemma 2.4, we have

$$\left|\widetilde{t}_n(X_i) - \widetilde{t}_{n,[i]}(X_i)\right| \leq O(1)\frac{a_n^2}{\rho_n n}. \tag{2.84}$$

So that by (2.84) and Schwarz inequality,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \widetilde{t}_{n,[i]}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big)\right|^2 \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \widetilde{t}_{n,[i]}(X_i)\big)^2 E_{n,\boldsymbol{\theta}}\big(t^*_{G_n}(X_i) - \theta_i\big)^2 \\
\leq\ & O(1)\frac{a_n^4}{\rho_n^2 n^2}.
\end{aligned}$$

Thus, the first term on the right hand side of (2.79) is bounded by

$$\left|\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \widetilde{t}_{n,[i]}(X_i)\big)\big(t^*_{G_n}(X_i) - \theta_i\big)\right| \leq O(1)\frac{a_n^2}{\rho_n n}. \tag{2.85}$$

Adding (2.76), 2.80, 2.81 and 2.85 together, we have

$$\frac{1}{n}\sum_{i=1}^{n} E_{n,\boldsymbol{\theta}}\big(\widetilde{t}_n(X_i) - \theta_i\big)^2 \leq 1 - J(\rho_n, G_n) + O(1)\frac{(\log n)^{3/2}}{\rho_n n}.$$

This completes the proof since $L_n(\widetilde{t}_n(\boldsymbol{X}), \boldsymbol{\theta}) = \sum_{i=1}^n \big(\widetilde{t}_n(X_i) - \theta_i\big)^2/n$. □

2.6 Some Simulation Results

2.6.1 Highlight of main results

Johnstone and Silverman [47] reported results of an extensive simulation study of 18 threshold estimators, including eight options of their EBThresh, the SURE and adaptive SURE [23], the FDR [1] at three levels, three block threshold methods [13, 14] and the soft and hard threshold at the universal threshold level $\sqrt{2\log n}$.

Table 2.1: Average total squared errors $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$ for $n = 1000$ unknown means in various binary models where θ_j is either 0 or μ with the number of nonzero $\theta_j = \mu$ being 5, 50 or 500. The "Best" stands for the best simulation results in Table 1 of Johnstone and Silverman. Each entry is based on 100 replications

# nonzero	5				50				500			
μ	3	4	5	7	3	4	5	7	3	4	5	7
James-Stein	45	76	113	199	314	448	560	715	820	892	929	964
EBThresh	**36**	**31**	**17**	**9**	**213**	**160**	**102**	72	858	876	788	661
SURE	42	62	72	74	417	609	211	211	832	837	838	838
FDR (.01)	44	50	24	**6**	392	302	128	**56**	2561	1334	667	527
FDR (.1)	41	34	**19**	14	279	173	113	101	1154	749	654	646
GMLEB	**39**	**32**	21	11	**161**	**112**	**58**	**15**	**461**	**291**	**133**	**20**
S-GMLEB	**33**	**26**	**16**	**6**	**154**	**106**	**53**	**10**	**457**	**288**	**130**	**17**
F-GEB	105	98	92	88	233	197	151	118	530	371	232	135
RF-GEB	72	64	58	56	203	154	102	77	**504**	**339**	**188**	**89**
HF-GEB	36	31	17	9	204	154	102	77	504	339	188	89
"Best"	34	32	17	5	201	156	95	52	829	730	609	505
Oracle	27	21	11	1	147	100	47	3	447	279	123	9

In their simulations, the overall best performer is the EBThresh using the posterior median for the prior (2.10) with the double exponential $dG_0(u)/du = e^{-|u|}/2$ and the MLE of (ω_0, τ).

In Table 2.1, we display our simulation results under exactly the same setting as in [47] for nine estimators: the James-Stein, the EBThresh [47] using the double exponential dG_0 in (2.10) and the MLE of (ω_0, τ), the SURE [23], the FDR [1] at levels q = 0.01 and q = 0.1, the GMLEB (2.15) with the uniform initialization, the S-GMLEB with the initialization (2.20), the F-GEB and HF-GEB as the Fourier general EB [74] and a hybrid [76] of its monotone version with the EBThresh. In each column, boldface entries denote the top three performers other than the hybrid estimator. We also display as "Best" the best of the simulation results in [47] over the 18 threshold estimators and as Oracle the average simulated risk of the oracle Bayes rule $t^*_{G_n}$ in (2.8).

These simulation results can be summarized as follows. The average ℓ_2 loss of the S-GMLEB happens to be the smallest among the nine estimators, with the S-GMLEB and GMLEB clearly outperforming all other methods by large margins for dense and moderately sparse signals. For very sparse signals, the S-GMLEB, the EBThresh, the GMLEB and the FDR estimators yield comparable results, and they all outperform the Fourier general EB and James-Stein estimators. Compared with the oracle, the regrets of the S-GMLEB and GMLEB are nearly fixed constants. Since the oracle prior (2.4) has a point mass at 0 in all the models used to generate data in this simulation experiment, the S-GMLEB yields slightly better results than the GMLEB as expected. The hybrid estimator correctly switches to the EBThresh for very sparse signals. These simulations and more presented in this subsection demonstrate the computational affordability of the proposed GMLEB. The most surprising aspect of the results in Table 2.1 is the strong performance of the both versions of the GMLEB for the most sparse signals with 0.5% of θ_i being nonzero, since the GMLEB is not specially designed to recover such signals (and threshold estimators are).

2.6.2 More simulation results

In addition to the simulation results reported in Table 2.1, we conducted more experiments to explore a larger sample size, sparse unknown means without exact zero, and iid unknown means from normal priors. The results for the nine statistical procedures and the oracle rule $t^*_{G_n}(\boldsymbol{X})$ for the general EB are reported in Tables 2.2-2.4, in the same format as Table 2.1. Each entry is based on an average of 100 replications. In each column, boldface entries indicate top three performers other than the hybrid estimator or the oracle.

In Table 2.2 we report simulation results for $n = 4000$. Compared with Table 2.1, F-GEB replaces EBThresh as a distant third top performer in the moderately sparse case of $\#\{i\colon \theta_i = \mu\} = 200$, and almost the same sets of estimators prevail as top performers in other columns. Since the collections of G_n are identical in Tables 2.1 and 2.2, the average squared loss $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2/n$ should decrease in n to

Table 2.2: Average total squared errors $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$ for $n = 4000$ unknown means in various binary models where θ_j is either 0 or μ with the number of nonzero $\theta_j = \mu$ being 5, 50 or 500

# nonzero	5				50				500			
μ	3	4	5	7	3	4	5	7	3	4	5	7
James-Stein	174	298	446	788	1245	1783	2228	2846	3274	3556	3704	3842
EBThresh	**140**	**120**	**68**	185	877	617	398	285	3426	3494	3135	2633
SURE	173	270	330	349	1729	824	826	826	3309	3329	3329	3329
FDR (.01)	175	203	106	**24**	1567	1221	509	227	10236	5370	2613	2095
FDR (.1)	160	135	76	54	1118	686	440	396	4634	2974	2605	2576
GMLEB	**136**	**115**	**71**	**31**	6**27**	**428**	**208**	**41**	**1818**	**1130**	**504**	**61**
S-GMLEB	**112**	**92**	**49**	**10**	**598**	**404**	**186**	**22**	**1801**	**1118**	**494**	**52**
F-GEB	237	220	180	163	724	546	349	218	1922	1250	643	250
RF-GEB	187	164	127	113	**686**	**496**	**278**	**149**	**1881**	**1202**	**585**	**174**
HF-GEB	140	120	68	185	686	496	278	149	1881	1202	585	174
Oracle	106	86	43	3	590	395	178	12	1778	1098	475	33

Table 2.3: Average of $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$: $n = 1000$, $\theta_j = \mu_j + \text{unif}[-0.2, 0.2]$, $\mu_j \in \{0, \mu\}$, $\#\{j \colon \mu_j = \mu\}$ =5, 50 or 500

# nonzero	5				50				500			
μ	3	4	5	7	3	4	5	7	3	4	5	7
James-Stein	57	87	123	206	318	449	558	711	819	889	925	959
EBThresh	**48**	**43**	**29**	**22**	**226**	**170**	**115**	86	859	875	789	665
SURE	55	74	84	86	428	623	219	219	832	837	837	837
FDR (.01)	56	60	37	**19**	397	319	141	**70**	2560	1357	668	533
FDR (.1)	52	48	32	27	289	189	127	114	1164	756	662	653
GMLEB	**48**	**42**	**31**	**22**	**171**	**123**	**68**	**25**	**466**	**300**	**144**	**32**
S-GMLEB	**43**	**38**	**28**	**19**	**165**	**119**	**65**	**23**	**463**	**297**	**142**	**30**
F-GEB	113	107	100	97	242	206	157	125	532	377	238	142
RF-GEB	81	74	67	65	213	167	111	85	**509**	**348**	**195**	**98**
HF-GEB	48	43	29	22	213	167	111	85	509	348	195	98
Oracle	39	34	23	14	160	114	60	16	455	290	135	23

Table 2.4: Average of $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$: $n = 1000$, iid $\theta_j \sim N(\mu, \sigma^2)$

σ^2	0.1				2				40			
μ	3	4	5	7	3	4	5	7	3	4	5	7
James-Stein	**93**	**92**	**93**	**93**	**671**	**672**	**671**	**673**	**981**	**981**	**981**	**980**
EBThresh	1089	1069	1047	1025	1023	1043	1040	1025	**994**	**996**	**997**	**1002**
SURE	1014	1486	3666	13530	995	1004	1041	3311	**993**	**995**	**996**	**1000**
FDR (.01)	3989	2073	1181	1005	2790	2233	1614	1066	1675	1614	1551	1433
FDR (.1)	1556	1105	1011	1004	1479	1259	1106	1010	1186	1173	1157	1120
GMLEB	**95**	**95**	**96**	**96**	**679**	**681**	**681**	**683**	1011	1012	1013	1010
S-GMLEB	**98**	**98**	**98**	**103**	**681**	**683**	**682**	**685**	1011	1012	1013	1010
F-GEB	171	169	171	175	745	747	744	750	1131	1120	1139	1125
RF-GEB	142	141	142	142	727	730	726	731	1083	1081	1086	1083
HF-GEB	142	141	142	142	727	730	726	731	1083	1081	1086	1083
Oracle	91	91	91	91	670	672	671	673	981	981	981	980

indicate convergence to the oracle risks for each estimator in each model, but this is not the case in entries in italics.

In Table 2.3, we report simulation results for sparse mean vectors without exact zero. It turns out that adding uniform $[0.2, 0.2]$ perturbations to θ_i does not change the results much, compared with Table 2.1.

In Table 2.4, we report simulation results for iid $\theta_i \sim N(\mu, \sigma^2)$. This is the parametric model in which the (oracle) Bayes estimators are linear. Indeed, the James-Stein estimator is the top performer throughout all the columns and tracks the oracle risk extremely well, while the GMLEB is not so far behind. It is interesting that for $\sigma^2 = 40$, the EBThresh and SURE outperform GMLEB as they approximate the naive $\widehat{\boldsymbol{\theta}} = \boldsymbol{X}$ with diminishing threshold levels. Another interesting phenomenon is the disappearance of the advantage of the S-GMLEB over the GMLEB, as the unknowns are no longer sparse.

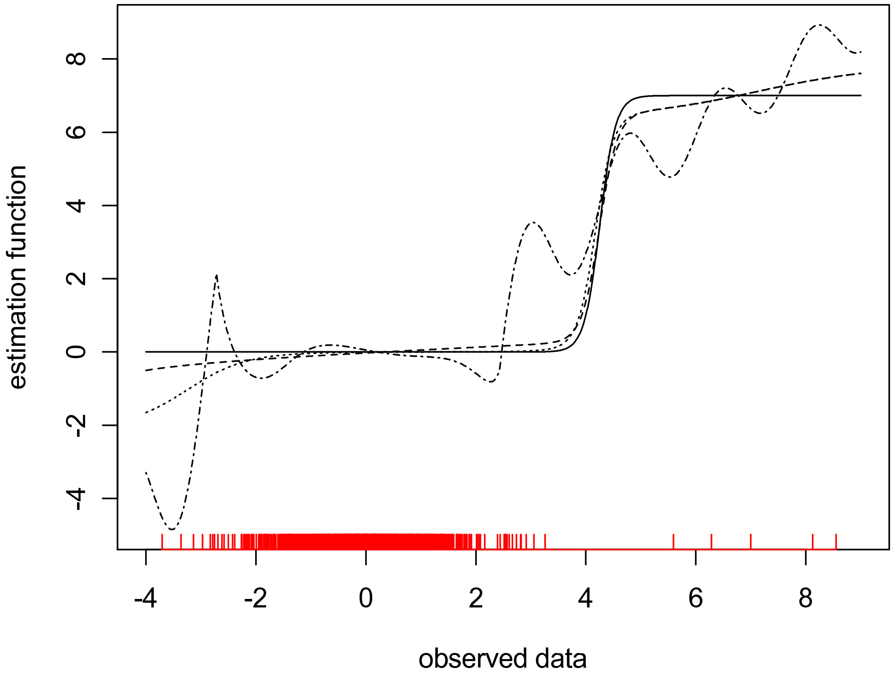

Figure 2.1: Plot of estimation functions of the Bayes estimator, GMLEB, S-GMLEB and F-GEB based on one set of data (1000 means, 5 nonzero means with $\mu = 7$). The solid, dashed and dotted curves are the estimation function of the Bayes estimator, the GMLEB and S-GMLEB respectively. The fluctuating curve represents the F-GEB. The upper tickmarks on the data axis present the data observed.

2.6.3 Additional simulations

The images of the Bayes estimator, GMLEB, S-GMLEB and F-GEB are shown in Figure 2.1. From Figure 2.1, we can see that around zero, the S-GMLEB is closer to the Bayes estimator than the GMLEB. The curve of the F-GEB is quite erratic, i.e., it is very data-dependent.

The reason why GEB-RML can improve over GEB-ML can be seen from Figure 2.2. GEB-RML estimates the prior $G_n(\theta)$ better than GEB-ML: GEB-RML puts more weights on 0 while GEB-ML puts lots of weights around 0.

2.7 Discussion

In this section, we discuss general EB with kernel estimates of the oracle Bayes rule, sure computation of an approximate generalized MLE and a number of

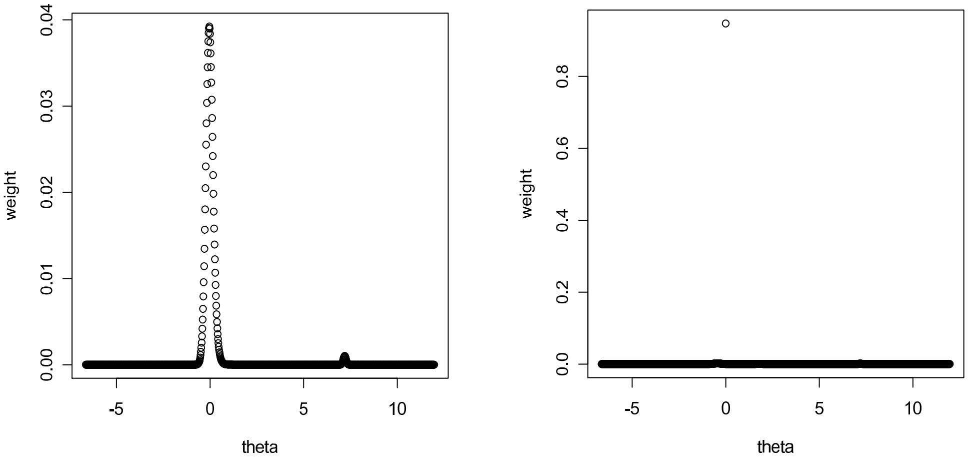

Figure 2.2: Scatter plot of weights of $\widehat{G}_n(\theta)$ of the GMLEB (left) and S-GMLEB (right) based on one set of data (1000 means, 5 nonzero means with $\mu = 7$).

additional issues.

2.7.1 Kernel methods

As discussed in Section 2.5, general EB estimators of the mean vector $\boldsymbol{\theta}$ can be directly derived from the formula (2.22) using the kernel method (2.66). This was done in [74] with the Fourier kernel $K(x, a) = \sin(ax)/(\pi x)$ and $\sqrt{2 \log n} \leq a_n \asymp \sqrt{\log n}$. The main rationale for using the Fourier kernel is the near optimal convergence rate of $\widehat{f}_n - f_{G_n} = O(\sqrt{(\log n)/n})$ and $\widehat{f}_n' - f_{G_n}' = O((\log n)/\sqrt{n})$, uniformly in $\boldsymbol{\theta}$. However, since the relationship between $\widehat{f}_n'(x)$ and $\widehat{f}_n(x)$ is not as trackable as in the case of generalized MLE $f_{\widehat{G}_n}$, a much higher regularization level $\rho_n \asymp \sqrt{(\log n)/n}$ in (2.66) were used [74, 76] to justify the theoretical results. This could be an explanation for the poor performance of the Fourier general EB estimator for very sparse $\boldsymbol{\theta}$ in our simulations. From this point of view, the GMLEB is much more appealing since its estimating function retains all analytic properties of the Bayes rule. Consequently, the GMLEB requires no regularization for the adaptive ratio optimality and adaptive minimaxity in our theorems.

Brown and Greenshtein [11] have studied (2.66) with the normal kernel $K(x) = \varphi(x)$ and possibly different bandwidth $1/a_n$, and have proved the adaptive ratio optimality (2.44) of their estimator when $\|\boldsymbol{\theta}\|_\infty$ and $R^*(G_{n,\boldsymbol{\theta}})$ have certain different polynomial orders. The estimating function $\widehat{t}_n(x)$ with the normal kernel, compared with the Fourier kernel, behaves more like the regularized Bayes rule (2.23) analytically with the positivity of $\widehat{f}_n(x)$ and more trackable relationship between $\widehat{f}_n'(x)$ and $\widehat{f}_n(x)$. Still, without some basic properties of the Bayes rule in Proposition 2.1 and Theorem 2.1, it is unclear if the kernel methods of the form (2.66) would possess as strong theoretical properties as in Theorems 2.3, 2.4, 2.5, 2.6 and 2.7 or perform as well as the GMLEB for moderate samples in simulations.

2.7.2 Sure computation of an approximate generalized MLE

We present a conservative data-driven criterion to guarantee (2.14) with the EM-algorithm. This provides a definitive way of computing the map from $\{X_i\}$ to $\widehat{G}_n$ in (2.14) and then to the GMLEB via (2.18).

Set $u_1 = \min_{1\le i\le n} X_i$, $u_m = \max_{1\le i\le n} X_i$, and

$$\epsilon = (u_m - u_1)/(m-1), \quad u_j = u_{j-1} + \epsilon. \tag{2.86}$$

Proposition 2.5. *Suppose $\epsilon^2\{(u_m-u_1)^2/4 + 1/8\} \le 1/n$ with a sufficient large m in (2.86). Let $\widehat{w}_j^{(0)} > 0\ \forall\ j \le m$ with $\sum_{j=1}^m \widehat{w}_j^{(0)} = 1$. Suppose that the EM-algorithm (2.19) is stopped at or beyond an iteration $k > 0$ with*

$$\max_{1\le j\le m} \log\Big(\widehat{w}_j^{(k)}/\widehat{w}_j^{(k-1)}\Big) \le \frac{1}{n}\log\Big(\frac{1}{eq_n}\Big), \tag{2.87}$$

where $q_n = (e\sqrt{2\pi}/n^2)\wedge 1$. Then, (2.14) holds for $\widehat{G}_n = \sum_{j=1}^m \widehat{w}_j^{(k)}\delta_{u_j}$.

Heuristically, smaller m provides larger $\min_j \widehat{w}_j^{(k)}$ and faster convergence of the EM algorithm, so that the "best choice" of m is

$$m - 2 < (u_m - u_1)\sqrt{n\{(u_m-u_1)^2/4 + 1/8\}} \le m - 1.$$

For $\max_i |X_i| \asymp \sqrt{\log n}$, this ensures the first condition of Proposition 2.5 with $m \asymp (\log n)\sqrt{n}$ and $\epsilon \asymp (n\log n)^{-1/2}$. Proposition 2.5 is proved via the smoothness of the normal density and Cover's upper bound [18, 71] for the maximum likelihood in finite mixture models.

2.7.3 Additional remarks

A crucial element for the theoretical results for the GMLEB is the oracle inequality for the regularized Bayes estimator with misspecified prior, as stated in Theorem 2.1. However, we do not believe that mathematical induction is sharp in the argument with higher and higher order of differentiation in the proof of Lemma 2.1. Consequently, the power κ_1 in Theorem 2.6 and 2.7 is larger than its counterpart more directly established for threshold estimators [1, 47]. Still, the GMLEB performs as well as any threshold estimators in our simulations for the most sparse mean vectors. As expected, the gain of the GMLEB is huge against the James-Stein estimator for sparse means and against threshold estimators for dense means.

It is interesting to observe in Table 2.1-2.3 that the simulated ℓ_2 risk for the GMLEB sometimes dips well below the benchmark $\sum_{i=1}^n \theta_i^2 \wedge 1 = \#\{i\colon \theta_i \neq 0\}$ for the oracle hard threshold rule $\widehat{\theta}_i = X_i I\{|\theta_i| \leq 1\}$ [36], while the simulated ℓ_2 risk for threshold estimators is always above that benchmark.

An important consequence of our results is the adaptive minimaxity and other optimality properties of the GMLEB approach to nonparametric regression under suitable smoothness conditions. For example, applications of the GMLEB estimator to the observed wavelet coefficients at individual resolution levels yield adaptive exact minimaxity in all Besov spaces as in [76].

The adaptive minimaxity (3.34) in Theorems 2.6 and 2.7 is uniform in the radii C for fixed shape p. A minimax theory for (weak) ℓ_p balls uniform in (p, C) can be developed by careful combination and improvement of the proofs in [21, 46, 76]. Since the oracle inequality (2.38) is uniform in p, uniform adaptive minimaxity in both p and C is in principle attainable for the GMLEB. The theoretical results in this chapter are all stated for deterministic $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. By either mild modifications of the proofs here or conditioning on the unknowns, analogues versions of our theorems can be established for the estimation of iid means $\{\xi_i\}$ in the EB model (2.48). Other possible directions of extension of the results in this chapter are the cases of $X_i \sim N(\theta_i, \sigma_n^2)$ via scale change, with known σ_n^2 or an independent consistent estimate of σ_n^2, and $X_i \sim N(\theta_i, \sigma_i^2)$ with known σ_i^2.

2.8 Proof

Here we prove Proposition 2.1, Lemma 2.1, Proposition 2.3, Theorem 2.3, 2.6 and 2.7 and then Proposition 2.5. We need one more lemma for the proof of Proposition 2.1. Throughout this subsection, $\lfloor x \rfloor$ denotes the greatest integer lower bound of x, and $\lceil x \rceil$ denotes the smallest integer upper bound of x.

Lemma 2.5. *Let $f_G(x)$ be as in (2.13) and $\widetilde{L}(y)$ as in Proposition 2.1. Then,*

$$\Big(\frac{f_G'(x)}{f_G(x)}\Big)^2 \le \frac{f_G''(x)}{f_G(x)} + 1 \le \widetilde{L}^2(f_G(x)) = \log\Big(\frac{1}{2\pi f_G^2(x)}\Big). \tag{2.88}$$

Proof of Lemma 2.5. Since $Y|\xi \sim N(\xi, 1)$ and $\xi \sim G$ under P_G, by (2.22)

$$\frac{f_G'(x)}{f_G(x)} = E_G[\xi - Y|Y = x], \quad \frac{f_G''(x)}{f_G(x)} + 1 = E_G[(\xi - Y)^2|Y = x].$$

This gives the first inequality of (2.88). The second inequality of (2.88) follows from Jensen's inequality: for $h(x) = e^{x/2}$,

$$h\Big(\frac{f_G''(x)}{f_G(x)} + 1\Big) \le E_G\Big[h\big((\xi - Y)^2\big)|Y = x\Big] = \frac{1}{\sqrt{2\pi} f_G(x)}.$$

This completes the proof. □

Proof of Proposition 2.1. Since $f_G(x) = \int \varphi(x-u)G(du) \ge 0$, the value of (2.23) is always between $t_G^*(x)$ and x. By Lemma 2.5

$$\big|t_G^*(x;\rho) - x\big| \le \frac{f_G(x)}{f_G(x) \vee \rho}\widetilde{L}\big(f_G(x)\big) \le \widetilde{L}(\rho)$$

for $\rho \le (2\pi e)^{-1/2}$, since $\widetilde{L}(y)$ is decreasing in y^2 and $y^2\widetilde{L}^2(y)$ is increasing in $y^2 \le 1/(2\pi e)$. Similarly, the second line of (2.24) follows from Lemma 2.5 and

$$\frac{\partial t_G^*(x;\rho)}{\partial x} = \begin{cases} 1 + f_G''(x)/f_G(x) - \{f_G'(x)/f_G(x)\}^2, & \text{if } f_G(x) > \rho, \\ 1 + f_G''(x)/\rho, & \text{if } f_G(x) < \rho. \end{cases}$$

Note that $\widetilde{L}(f_G(x)) \le \widetilde{L}(\rho)$ for $f_G(x) \ge \rho$, and for $f_G(x) < \rho \le (2\pi e^3)^{-1/2}$,

$$0 \le 1 - \frac{f_G(x)}{\rho} \le 1 + \frac{f_G''(x)}{\rho} \le 1 + \frac{f_G(x)}{\rho}\Big(\widetilde{L}^2(f_G(x)) - 1\Big) \le \widetilde{L}^2(\rho)$$

due to the monotonicity of $y\{\widetilde{L}^2(y) - 1\}$ in $0 \le y \le (2\pi e^3)^{-1/2}$. □

Proof of Lemma 2.1. Let $D = d/dx$. We first prove that for all integers $k \ge 0$ and $a \ge \sqrt{2k-1}$,

$$\int \{D^k(f_G - f_{G_0})\}^2 dx \le \frac{4a^{2k}}{\sqrt{2\pi}} d^2(f_G, f_{G_0}) + \frac{4a^{2k-1}}{\pi} e^{-a^2}. \tag{2.89}$$

Let $h^*(u) = \int e^{iux} h(x)dx$ for all integrable h. Since $|f_G^*(u)| \le \varphi^*(u) = e^{-u^2/2}$, it follows from the Plancherel identity that

$$\begin{aligned} \int \{D^k(f_G - f_{G_0})\}^2 dx &= \frac{1}{2\pi}\int u^{2k}|f_G^*(u) - f_{G_0}^*(u)|^2 du \\ &\le \frac{a^{2k}}{2\pi}\int |f_G^*(u) - f_{G_0}^*(u)|^2 du + \frac{4}{2\pi}\int_{|u|>a} u^{2k}e^{-u^2}du \\ &= a^{2k}\int |f_G - f_{G_0}|^2 dx + \frac{4}{\pi}c_k, \end{aligned}$$

where $c_k = \int_{u>a} u^{2k}e^{-u^2}du$. Since $(k-1/2) \le a^2/2$, integrating by parts yields

$$\begin{aligned} c_k &= 2^{-1}a^{2k-1}e^{-a^2} + \{(k-1/2)/a^2\}a^2 c_{k-1} \\ &\le 2^{-1}a^{2k-1}e^{-a^2}(1 + 1/2 + \cdots + 1/2^{k-1}) + 2^{-k}a^{2k}c_0 \\ &\le a^{2k-1}e^{-a^2} \end{aligned}$$

due to $c_0 \le a^{-1}\int_{u>a} ue^{-u^2}du = e^{-a^2}/(2a)$. Since $f_G(x) \le 1/\sqrt{2\pi}$,

$$\int |f_G - f_{G_0}|^2 dx \le \Big\|\sqrt{f_G} + \sqrt{f_{G_0}}\Big\|_\infty^2 d^2(f_G, f_{G_0}) \le \frac{4}{\sqrt{2\pi}} d^2(f_G, f_{G_0}).$$

The combination of the above inequalities yields (2.89).

Define $w_* = 1/(f_G \vee \rho + f_{G_0} \vee \rho)$ and $\Delta_k = (\int \{D^k(f_G - f_{G_0})\}^2 w_*)^{1/2}$. Integrating by parts, we find

$$\Delta_k^2 = -\int \{D^{k-1}(f_G - f_{G_0})\}\{D^{k+1}(f_G - f_{G_0})w_* + (D^k(f_G - f_{G_0}))(Dw_*)\}.$$

Since $|(Dw_*)(x)| \le 2\widetilde{L}(\rho)w_*(x)$ by Proposition 2.1, Cauchy-Schwarz gives

$$\Delta_k^2 \le \Delta_{k-1}\Delta_{k+1} + 2\widetilde{L}(\rho)\Delta_{k-1}\Delta_k.$$

Let k_0 be a nonnegative integer satisfying $k_0 \le \widetilde{L}^2(\rho)/2 < k_0+1$. Define $k^* = \min\{k\colon \Delta_{k+1} \le k_0 2\widetilde{L}(\rho)\Delta_k\}$. For $k < k^*$, we have $\Delta_k^2 \le (1+1/k_0)\Delta_{k-1}\Delta_{k+1}$, so that for $k^* \le k_0$,

$$\frac{\Delta_1}{\Delta_0} \le \Big(1+\frac{1}{k_0}\Big)\frac{\Delta_2}{\Delta_1} \le \Big(1+\frac{1}{k_0}\Big)^{k^*}\frac{\Delta_{k^*+1}}{\Delta_{k^*}} \le ek_0 2\widetilde{L}(\rho) \le e\widetilde{L}^3(\rho).$$

Since $\big(f_G^{1/2}+f_{G_0}^{1/2}\big)^2 w_* \le 2$, we have $\Delta_0^2 \le 2d^2(f_G, f_{G_0})$. Thus, for $k^* \le k_0$,

$$\Delta_1 \le e\widetilde{L}^3(\rho)\sqrt{2}d(f_G, f_{G_0}). \tag{2.90}$$

For $k_0 < k^*$, $\Delta_1/\Delta_0 \le (1+1/k_0)^k \Delta_{k+1}/\Delta_k$ for all $k \le k_0$, so that

$$\begin{aligned} \frac{\Delta_1}{\Delta_0} &\le \Big[\prod_{k=0}^{k_0}\Big\{(1+1/k_0)^k\Delta_{k+1}/\Delta_k\Big\}\Big]^{1/(k_0+1)} \\ &= (1+1/k_0)^{k_0/2}\Big\{\Delta_{k_0+1}/\Delta_0\Big\}^{1/(k_0+1)}. \end{aligned} \tag{2.91}$$

To bound Δ_{k_0+1} by (2.89), we pick the constant $a > 0$ with the a^2 in (2.29), so that $a^2 \ge 2(k_0+1/2)$ and $e^{-a^2} \le d^2(f_G, f_{G_0})$. Since $w_* \le 1/(2\rho)$, an application of (2.89) with this a gives

$$\begin{aligned} \Delta_{k_0+1}^2 &\le \frac{1}{2\rho}\int \{D^{k_0+1}(f_G - f_{G_0})\}^2 \\ &\le \frac{2a^{2(k_0+1)}}{\rho\sqrt{2\pi}} d^2(f_G, f_{G_0})\big(1+a^{-1}\sqrt{2/\pi}\big). \end{aligned}$$

Since $\Delta_0^2 \le 2d^2(f_G, f_{G_0})$, inserting the above inequality into (2.91) yields

$$\begin{aligned} \Delta_1 &\le (1+1/k_0)^{k_0/2}\Delta_0^{k_0/(k_0+1)}\Delta_{k_0+1}^{1/(k_0+1)} \\ &\le (1+1/k_0)^{k_0/2}\sqrt{2}d(f_G, f_{G_0})a\Big(\frac{1+\sqrt{2/\pi}}{\rho\sqrt{2\pi}}\Big)^{1/(2k_0+2)} \\ &\le \sqrt{e}d(f_G, f_{G_0})a\sqrt{2}(2\pi\rho^2)^{-1/(4k_0+4)}. \end{aligned} \tag{2.92}$$

Since $|\log(2\pi\rho^2)| = \widetilde{L}^2(\rho) < 2k_0+2$, (2.29) follows from (2.90) and (2.92). □

Proof of Proposition 2.3. We provide a dense version of the proof since it is similar to the entropy calculation in [39, 40, 80].

It follows from (2.23), (2.24) and Lemma 2.5 that

$$\Big|t_G^*(x;\rho) - t_H^*(x;\rho)\Big| \le \frac{1}{\rho}\Big|f_G'(x) - f_H'(x)\Big| + \frac{\widetilde{L}(\rho)}{\rho}\Big|f_G(x) - f_H(x)\Big|, \tag{2.93}$$

so that we need to control the norm of both f_G and f'_G.

Let $a = \widetilde{L}(\eta)$, $j^* = \lceil 2M/a + 2 \rceil$ and $k^* = \lfloor 6a^2 \rfloor$. Define semiclosed intervals

$$I_j = \big(-M + (j-2)a, (-M + (j-1)a) \wedge (M+a) \big],\ j = 1, \ldots, j^*,$$

to form a partition of $(-M-a, M+a]$. It follows from the Carathéodory's theorem [17] that for each distribution function G there exists a discrete distribution function G_m with support $[-M-a, M+a]$ and no more than $m = (2k^*+2)j^*+1$ support points such that

$$\int_{I_j} u^k G(du) = \int_{I_j} u^k G_m(du),\ k = 0, 1, \ldots, 2k^*+1,\ j = 1, \ldots, j^*. \tag{2.94}$$

Since the Taylor expansion of $e^{-t^2/2}$ has alternating signs, for $t^2/2 \le k^* + 2$,

$$0 \le \mathrm{Rem}(t) = (-1)^{k^*+1} \Big\{ \varphi(t) - \sum_{k=0}^{k^*} \frac{(-t^2/2)^k}{k!\sqrt{2\pi}} \Big\} \le \frac{(t^2/2)^{k^*+1}}{(k^*+1)!\sqrt{2\pi}}.$$

Thus, since $k^* + 1 \ge 6a^2$, for $x \in I_j \cap [-M, M]$, the Stirling formula yields

$$\begin{aligned} & \Big| f'_G(x) - f'_{G_m}(x) \Big| \\ \le\ & \Big| \int_{(I_{j-1} \cup I_j \cup I_{j+1})^c} (x-u)\varphi(x-u)\{G(du) - G_m(du)\} \Big| \\ & + \Big| \int_{I_{j-1} \cup I_j \cup I_{j+1}} (x-u)\mathrm{Rem}(x-u)\{G(du) - G_m(du)\} \Big| \\ \le\ & \max_{t \ge a} t\varphi(t) + \frac{4a\{(2a)^2/2\}^{k^*+1}}{\sqrt{2\pi}(k^*+1)!} \le a\eta + \frac{4a(e/3)^{k^*+1}}{2\pi(k^*+1)^{1/2}} \end{aligned} \tag{2.95}$$

due to $a \ge 1$. Similarly, for $|x| \le M$,

$$\Big| f_G(x) - f_{G_m}(x) \Big| \le \eta + \frac{(e/3)^{k^*+1}}{2\pi(k^*+1)^{1/2}}. \tag{2.96}$$

Furthermore, since $(e/3)^6 \le e^{-1/2}$ and $k^* + 1 \ge 6a^2 \ge 6$, we have $(e/3)^{k^*+1} \le e^{-a^2/2} = \sqrt{2\pi}\eta$, so that by (2.93), (2.95) and (2.96),

$$\begin{aligned} & \Big\| t^*_G(\cdot; \rho) - t^*_{G_m}(\cdot; \rho) \Big\|_{\infty, M} \\ \le\ & \rho^{-1} \Big(a\eta + \frac{4ae^{-a^2/2}}{2\pi\sqrt{6a^2}} \Big) + \rho^{-1} \widetilde{L}(\rho) \Big(\eta + \frac{e^{-a^2/2}}{2\pi\sqrt{6a^2}} \Big) \\ \le\ & \rho^{-1}\eta \Big(2\widetilde{L}(\eta) + 5/\sqrt{12\pi} \Big). \end{aligned} \tag{2.97}$$

Let $\xi \sim G_m$, $\xi_\eta = \eta\,\mathrm{sgn}(\xi)\lfloor|\xi|/\eta\rfloor$ and $G_{m,\eta} \sim \xi_\eta$. Since $|\xi - \xi_\eta| \le \eta$,

$$\Big\|f_{G_m} - f_{G_{m,\eta}}\Big\|_\infty \le C_1^*\eta, \quad \Big\|f'_{G_m} - f'_{G_{m,\eta}}\Big\|_\infty \le C_2^*\eta,$$

where $C^* = \sup_x|\varphi'(x)| = (2e\pi)^{-1/2}$ and $C_2^* = \sup_x|\varphi''(x)| = \sqrt{2/\pi}e^{-3/2}$. This and (2.93) imply

$$\Big\|t^*_{G_m}(\cdot;\rho) - t^*_{G_{m,\eta}}(\cdot;\rho)\Big\|_\infty \le \frac{\eta}{\rho}\Big\{C_2^* + C_1^*\widetilde{L}(\rho)\Big\}. \tag{2.98}$$

Moreover, $G_{m,\eta}$ has at most m support points.

Let $\mathscr{P}^m$ be the set of all vectors $\boldsymbol{w} = (w_1, \ldots, w_m)$ satisfying $w_j \ge 0$ and $\sum_{j=1}^m w_j = 1$. Let $\mathscr{P}^{m,\eta}$ be an η-net of $N(\eta, \mathscr{P}^m, \|\cdot\|_1)$ elements in $\mathscr{P}^m$:

$$\inf_{\boldsymbol{w}^{m,\eta}\in\mathscr{P}^{m,\eta}} \|\boldsymbol{w} - \boldsymbol{w}^{m,\eta}\|_1 \le \eta, \quad \forall\, \boldsymbol{w} \in \mathscr{P}^m.$$

Let $\{u_j, j = 1, \ldots, m\}$ be the support of $G_{m,\eta}$ and $\boldsymbol{w}^{m,\eta}$ be a vector in $\mathscr{P}^{m,\eta}$ with $\sum_{j=1}^m |G_{m,\eta}(\{u_j\}) - w_j^{m,\eta}| \le \eta$. Set $\widetilde{G}_{m,\eta} = \sum_{j=1}^m w_j^{m,\eta}\delta_{u_j}$. Then,

$$\Big\|f_{G_{m,\eta}} - f_{\widetilde{G}_{m,\eta}}\Big\|_\infty \le C_0^*\eta, \quad \Big\|f'_{G_{m,\eta}} - f'_{\widetilde{G}_{m,\eta}}\Big\|_\infty \le C_1^*\eta,$$

where $C_0^* = \varphi(0) = 1/\sqrt{2\pi}$. This and (2.93) imply

$$\Big\|t^*_{G_{m,\eta}}(\cdot;\rho) - t^*_{\widetilde{G}_{m,\eta}}(\cdot;\rho)\Big\|_\infty \le \frac{\eta}{\rho}\Big\{C_1^* + C_0^*\widetilde{L}(\rho)\Big\}. \tag{2.99}$$

The support of $G_{m,\eta}$ and $\widetilde{G}_{m,\eta}$ is $\Omega_{\eta,M} = \{0, \pm\eta, \pm 2\eta, \ldots\} \cap [-M-a, M+a]$.

Summing (2.97), (2.98) and (2.99) together, we find

$$\begin{aligned} &\Big\|t^*_G(\cdot;\rho) - t^*_{\widetilde{G}_{m,\eta}}(\cdot;\rho)\Big\|_{\infty,M} \\ \le\ & (\eta/\rho)\Big[\{2 + C_1^* + C_0^*\}\widetilde{L}(\eta) + 5/\sqrt{12\pi} + C_2^* + C_1^*\Big] \\ \le\ & (\eta/\rho)\Big\{2.65\widetilde{L}(\eta) + 1.24\Big\} \le \eta^*. \end{aligned}$$

Counting the number of ways to realize $\{u_j\}$ and $\boldsymbol{w}^{m,\eta}$, we find

$$N(\eta^*, \mathscr{T}_\rho, \|\cdot\|_{\infty,M}) \le \binom{|\Omega_{\eta,M}|}{m} N(\eta, \mathscr{P}^m, \|\cdot\|_1), \tag{2.100}$$

with $m = (2k^* + 2)j^* + 1$, $|\Omega_{\eta,M}| = 1 + 2\lfloor(M+a)/\eta\rfloor$, $a = \widetilde{L}(\eta)$, $j^* = \lceil 2M/a + 2\rceil$ and $k^* = \lfloor 6a^2\rfloor$.

Since $\mathscr{P}^m$ is in the ℓ_1 unit-sphere of $\mathbb{R}^m$, $N(\eta, \mathscr{P}^m, \|\cdot\|_1)$ is no greater than the maximum number of disjoint $\mathrm{Ball}(\boldsymbol{v}_j, \eta/2, \|\cdot\|_1)$ with centers $\boldsymbol{v}_j$ in the unit sphere. Since all these balls are inside the $(1+\eta/2)$ ℓ_1-ball, volume comparison yields $N(\eta, \mathscr{P}^m, \|\cdot\|_1) \leq (2/\eta+1)^m$. With another application of the Stirling formula, this and (2.100) yield

$$\begin{aligned} & N(\eta^*, \mathscr{T}_\rho, \|\cdot\|_{\infty,M}) \\ \leq\ & (2/\eta+1)^m |\Omega_{\eta,M}|^m/m! \\ \leq\ & \{(1+2/\eta)(1+2(M+a)/\eta)\}^m \{(m+1)^{m+1/2}e^{-m-1}\sqrt{2\pi}\}^{-1} \\ \leq\ & \{(\eta+2)(\eta+2(M+a))e/(m+1)\}^m \eta^{-2m} e\{2\pi(m+1)\}^{-1/2}. \end{aligned} \tag{2.101}$$

Since $m-1 \geq 12a^2(2M/a+2) = 24a(M+a)$ and $a \geq 1 \geq 1/2 \geq \eta$,

$$(\eta+2)(\eta+2(M+a))e \leq 8\{1/2+2(M+a)\} \leq m+1.$$

Hence, (2.104) is bounded by η^{-2m} with $m \leq 2(6a^2+1)(2M/a+3)+1$. □

Proof of Theorem 2.3. Throughout the proof, we use M_0 to denote a universal constant which may take different values from one appearance to another. For simplicity, we take $q_n = (e\sqrt{2\pi}/n^2) \wedge 1$ in (2.14) so that (2.34) holds with $\rho_n = n^{-3}$.

Let ϵ_n and x_* be as in Theorem 2.2 and $\widetilde{L}(\rho) = \sqrt{-\log(2\pi\rho^2)}$ be as in Proposition 2.1 and 2.4. With $\rho_n = n^{-3}$, set

$$\eta = \frac{\rho_n}{n} = \frac{1}{n^4},\ \eta^* = \frac{\eta}{\rho_n}\{3\widetilde{L}(\eta)+2\},\ M = \frac{2n\epsilon_n^2}{(\log n)^{3/2}}. \tag{2.102}$$

Let $x^* = \max(x_*, 1)$ and $\{t^*_{H_j}(\cdot;\rho_n), j \leq N\}$ ba a $(2\eta^*)$-net of

$$\mathscr{T}_{\rho_n} \cap \{t^*_G\colon d(f_G, f_{G_n}) \leq x^*\epsilon_n\} \tag{2.103}$$

under the $\|\cdot\|_{\infty,M}$ seminorm as in proposition 2.3, with distributions H_j satisfying $d(f_{H_j}, f_{G_n}) \leq x^*\epsilon_n$ and $N = N(\eta^*, \mathscr{T}_{\rho_n}, \|\cdot\|_{\infty,M})$. It is a $(2\eta^*)$-net due to the additional requirement on H_j. Since $M \geq 4\sqrt{\log n}$ and $\eta = 1/n^4$ by (2.32) and (2.102), Proposition 2.3 and (2.102) give

$$\log N \leq M_0(\log n)^{3/2}M/2 \leq M_0 n\epsilon_n^2. \tag{2.104}$$

We divide the ℓ_2 distance of the error into 5 parts:

$$\Big\| t^*_{\widehat{G}_n}(\boldsymbol{X};\rho_n) - \boldsymbol{\theta} \Big\| \le \sqrt{nR^*(G_n)} + \sum_{j=1}^4 \zeta_{jn},$$

where ζ_{jn} are as in (2.39), (2.40), (2.41) and (2.42). As we have mentioned in the outline, the problem is to bound $E_{n,\boldsymbol{\theta}}\zeta_{jn}^2$ in view of (2.43).

Let A_n and ζ_{1n} be as in (2.39). Since $x^* = 1 \vee x_* \ge 1$ and $n\epsilon^2 \ge 2(\log n)^2$ by (2.32), Theorem 2.2 gives $P_{n,\boldsymbol{\theta}}\{A_n^c\} \le \exp\big(-(x^*)^2 n\epsilon_n^2/(2\log n)\big) \le 1/n$. Thus, since $\widetilde{L}^2(\rho_n) = -\log(2\pi/n^6)$ with $\rho_n = n^{-3}$, Proposition 2.1 gives

$$\begin{aligned} E_{n,\boldsymbol{\theta}}\zeta_{1n}^2 &= E_{n,\boldsymbol{\theta}} \sum_{i=1}^n \Big\{ (t^*_{\widehat{G}}(X_i;\rho_n) - X_i) + (X_i - \theta_i) \Big\}^2 I_{A_n^c} \\ &\le 2n\widetilde{L}^2(\rho_n) P_{n,\boldsymbol{\theta}}\{A_n^c\} + 2E_{n,\boldsymbol{\theta}} \sum_{i=1}^n (X_i - \theta_i)^2 I_{A_n^c} \\ &\le M_0 \log n + 2n \int_0^\infty \min\big(P\{|N(0,1)| > x\}, 1/n\big) dx^2. \end{aligned}$$

Since $P\{N(0,1) > x\} \le e^{-x^2/2}$ and $\int_0^\infty \min(ne^{-x^2/2}, 1) dx^2/2 = 1 + \log n$,

$$E_{n,\boldsymbol{\theta}}\zeta_{1n}^2 \le M_0 \log n \le M_0 n\epsilon_n^2. \tag{2.105}$$

Consider ζ_{2n}^2. Since $t^*_{H_j}(\cdot;\rho_n)$ form a $(2\eta^*)$-net of (2.103) under $\|\cdot\|_{\infty,M}$ and $|t^*_G(x;\rho) - x| \le \widetilde{L}(\rho)$ by Proposition 2.1, it follows from (2.40) that

$$\begin{aligned} \zeta_{2n}^2 &\le \min_{j \le N} \|t^*_{\widehat{G}_n}(\boldsymbol{X};\rho_n) - t^*_{H_j}(\boldsymbol{X};\rho_n)\|^2 I_{A_n} \\ &\le (2\eta^*)^2 \#\{i\colon |X_i| \le M\} + \{2\widetilde{L}(\rho_n)\}^2 \#\{i\colon |X_i| > M\}. \end{aligned}$$

By (2.32), $(n\epsilon_n^2/\log n)^{p+1} \ge n\{\sqrt{\log n}\,\mu_p^w(G_n)\}^p$, so that by (2.31) and (2.102),

$$\int_{|u|\ge M/2} G_n(du) \le \Big(\frac{\mu_p^w(G_n)}{M/2}\Big)^p \le \Big(\frac{2n\epsilon_n^2}{M(\log n)^{3/2}}\Big)^p \frac{\epsilon_n^2}{\log n} = \frac{\epsilon_n^2}{\log n}. \tag{2.106}$$

Thus, since $\eta^* = n^{-1}\{3\widetilde{L}(n^{-4}) + 2\}$ and $M \ge 4\sqrt{\log n}$ by (2.102) and (2.32),

$$\begin{aligned} E_{n,\boldsymbol{\theta}}\zeta_{2n}^2 &\le n(2\eta^*)^2 + 4\widetilde{L}^2(n^{-3}) E_{n,\boldsymbol{\theta}} \#\{i\colon |X_i| > M\} \\ &\le M_0(\log n) n \Big(\frac{1}{n^2} + \int_{|u|\ge M/2} G_n(du) + P\{|N(0,1)| > 2\sqrt{\log n}\} \Big) \\ &\le M_0(\log n) \Big(\frac{1}{n} + \frac{n\epsilon_n^2}{\log n} + \frac{2}{n} \Big). \end{aligned}$$

Since $n\epsilon_n^2 \geq 2(\log n)^2$ by (2.32), we find

$$E_{n,\boldsymbol{\theta}}\zeta_{2n}^2 \leq M_0 n\epsilon_n^2. \tag{2.107}$$

Now, consider ζ_{3n}^2. Since $\widetilde{L}^2(\rho_n) \leq M_0 \log n$, it follows from (2.41), Proposition 2.4 and (2.104) that

$$\begin{aligned} E_{n,\boldsymbol{\theta}}\zeta_{3n}^2 &= \int_0^\infty P_{n,\boldsymbol{\theta}}\{\zeta_{3n} > x\}dx^2 \\ &\leq \int_0^\infty \min\Big\{1, N\exp\big(-x^2/(2\widetilde{L}^4(\rho_n))\Big\}dx^2 \\ &= 2\widetilde{L}^4(\rho_n)(1+\log N) \\ &\leq M_0(\log n)^2 n\epsilon_n^2. \end{aligned} \tag{2.108}$$

For ζ_{4n}^2, it suffices to apply Theorem 2.1 (ii) with $G_0 = G_n$, $G = H_j$, $\rho = \rho_n = n^{-3}$, $x_0 = M/2$ and $\epsilon_0 = x^*\epsilon_n \geq d(f_{H_j}, f_{G_n})$, since

$$\zeta_{4n}^2 \leq n \max_{j\leq N}\Big\{E_{G_n}\{t_{H_j}^*(Y;\rho_n) - \xi)^2 - R^*(G_n)\Big\} \tag{2.109}$$

by (2.42) and (2.5). It follows from (2.106) that the M_1 in Theorem 2.1 (ii) is no greater than

$$\frac{\int_{|u|\geq M/2} G_n(du)}{|\log\rho_n|^3(x^*\epsilon_n)^2} \leq \frac{\epsilon_n^2/\log n}{(3\log n)^3\epsilon_n^2} \leq M_0.$$

Since $M = 2n\epsilon_n^2/(\log n)^{3/2}$ by (2.102) and $n\epsilon_n^2 \geq 2(\log n)^2$ by (2.32), the M_2 in Theorem 2.1 (ii) is no greater than

$$\frac{2(M/2+1)\rho_n}{(\log\rho_n)^2(x^*\epsilon_n)^2} \leq \frac{2(n\epsilon_n^2/(\log n)^{3/2}+1)/n^3}{(3\log n)^2\epsilon_n^2} \leq \frac{\sqrt{\log n}+1}{n^2(\log n)^4} \leq M_0$$

with $\rho_n = n^{-3}$. Thus, by Theorem 2.1 (ii) and (2.109)

$$\zeta_{4n}^2 \leq M_0 n|(\log\rho_n)/3|^3\epsilon_n^2 = M_0 n\epsilon_n^2(\log n)^3. \tag{2.110}$$

Adding (2.105), (2.107), (2.108) and (2.109) together, we have

$$E_{n,\boldsymbol{\theta}}\Big(\sum_{j=1}^4 |\zeta_{jn}|\Big)^2 \leq M_0 n\epsilon_n^2(\log n)^3.$$

Since $L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2/n$, this and (2.43) complete the proof. □

Proof of Theorem 2.6. As we have mentioned, by (2.59), (2.60) and (2.61), the adaptive minimaxity (3.34) with $\Theta_n = \Theta_{p,C_n,n}$ follows from (2.62). By (2.31) and (2.54), $\mu_p^w(G_{n,\boldsymbol{\theta}}) \le C$ for $\boldsymbol{\theta} \in \Theta_{p,C,n}$, so that by (2.32) and Theorem 2.3, $\sup_{\boldsymbol{\theta}\in\Theta_{p,C,n}} \widetilde{r}_{n,\boldsymbol{\theta}}(\widehat{t}_n) \le \epsilon_{p,C,n}(\log n)^{3/2}$ with

$$\epsilon^2_{p,C,n} = \max\Big[2\log n, \big\{nC^p(\log n)^{p/2}\big\}^{1/(1+p)}\Big](\log n)/n. \tag{2.111}$$

Thus, it suffices to verify that for sequences C_n satisfying (2.55),

$$\epsilon^2_{p,C_n,n}(\log n)^3/J_{p,C_n} \to 0, \tag{2.112}$$

where $J_{p,C} = \min\{1, C^{p\wedge 2}\{1 \vee (2\log(1/C^p))\}^{(1-p/2)_+}\}$ as in (2.62).

We consider three cases. For $C_n^{2\wedge p} > e^{-1/2}$, $J_{p,C_n} \ge e^{-1/2}$ and

$$\epsilon^2_{p,C_n,n}(\log n)^3 = \max\Big[\frac{2(\log n)^5}{n}, \big\{C_n(\log n)^{9/2+4/p}/n\big\}^{p/(1+p)}\Big] = o(1),$$

since $\kappa_2(p) = 9/2 + 4/p$ in (2.55).

For $p < 2$ and $C_n^p \le e^{-1/2}$, $J_{p,C_n} = C_n^p\{2\log n(1/C_n^p)\}^{1-p/2}$, so that by (2.112),

$$\begin{aligned} &\epsilon^2_{p,C_n,n}(\log n)^3/J_{p,C_n} \\ = \;&\max\Big[\frac{2(\log n)^5}{nC_n^p\{2\log(1/C_n^p)\}^{1-p/2}}, \frac{(\log n)^{4+p/(2+2p)}}{(nC_n^p)^{p/(1+p)}\{2\log(1/C_n^p)\}^{1-p/2}}\Big]. \end{aligned}$$

Since the case $C_n^p > n^{-1/2}$ is trivial, it suffices to consider the case $C_n^p \le n^{-1/2}$ where

$$\frac{\epsilon^2_{p,C_n,n}(\log n)^3}{J_{p,C_n}} \asymp \max\Big[\frac{(\log n)^{4+p/2}}{nC_n^p}, \frac{(\log n)^{3+p/2+p/(2+2p)}}{(nC_n^p)^{p/(1+p)}}\Big].$$

Since $4 + p/2 \le p\kappa_1(p) = 4 + 3/p + p/2 = (1 + 1/p)\{3 + p/2 + p/(2 + 2p)\}$, (2.55) still implies (2.112).

Finally, for $p \ge 2$ and $C_n^2 \le e^{-1/2}$, $J_{p,C_n} = C_n^2$, so that

$$\frac{\epsilon^2_{p,C_n,n}(\log n)^3}{J_{p,C_n}} = \max\Big[\frac{2(\log n)^5}{nC_n^2}, \Big\{\frac{C_n(\log n)^{9/2+4/p}}{nC_n^{2(1+1/p)}}\Big\}^{p/(1+p)}\Big].$$

Since $nC_n^{1+2/p} = n^{1/2-1/p}(nC_n^2)^{1/2+1/p}$, we need $(\log n)^5/(nC_p^2) \to 0$ for $p > 2$ and $(\log n)^{13/2}/(nC_n^2) \to 0$ for $p = 2$. Again (2.55) implies (2.112). □

Proof of Theorem 2.7. Since the oracle inequality (2.38) is based on the weak ℓ_p norm, the proof of Theorem 2.6 also provides (2.65). $\square$

Proof of Proposition 2.5. Let $\widehat{G}_n^*$ be the exact generalized MLE as in (2.12). Since $\varphi(x)$ is decreasing in $|x|$, we have $\widehat{G}_n^*([u_1, u_m]) = 1$. Let $I_j = (u_{j-1}, u_j]$ and $I_j^* = [u_{j-1}, u_j]$ for $j \geq 2$ and $I_1 = I_1^* = \{u_1\}$. Let $H_{m,j}$ be sub-distributions with support $\{u_{j-1}, u_j\} \cap I_j^*$ such that

$$H_{m,j}(I_j^*) = \widehat{G}_n^*(I_j), \ \int_{I_j^*} u H_{m,j}(du) = \int_{I_j} u \widehat{G}_n^*(du), \ 1 \leq j \leq m. \tag{2.113}$$

Let $j > 1$ and $x \in [u_1, u_m]$ be fixed. Set $x_j = x - (u_j + u_{j-1})/2$ and $t = u - (u_j + u_{j-1})/2$ for $u \in I_j^*$. Since $|x_j t| \leq (u_m - u_1)\epsilon/2 \leq n^{-1/2} \leq 1$,

$$-(1 - e^{-t^2/2}) e^{x_j t} \leq e^{x_j t - t^2/2} - (1 + x_j t) \leq x_j^2 t^2 e^{x_j t - t^2/2}, \tag{2.114}$$

where the second inequality follows from $e^{-t^2/2}(1 - x_j t) \leq e^{-x_j t}$. Since $\varphi(x-u) = \varphi(x_j - t) = \varphi(x_j)\exp(x_j t - t^2/2)$, (2.113) and (2.114) yield

$$\begin{aligned}
& \int_{I_j} \varphi(x-u)\widehat{G}_n^*(du) - \int_{I_j^*} \varphi(x-u) H_{m,j}(du) \\
\leq \ & \int_{I_j} x_j^2 t^2 \varphi(x-u) \widehat{G}_n^*(du) + \int_{I_j^*} (e^{t^2/2} - 1)\varphi(x-u) H_{m,j}(du) \\
\leq \ & (u_m - u_1)^2 (\epsilon/2)^2 \int_{I_j} \varphi(x-u)\widehat{G}_n^*(du) + (e^{\epsilon^2/8} - 1) \int_{I_j^*} \varphi(x-u) H_{m,j}(du).
\end{aligned}$$

Let $H_m = \sum_{j=1}^m H_{m,j}$. Summing the above inequality over j, we find $e^{\epsilon^2/8} f_{H_m}(x) \geq (1-\eta) f_{\widehat{G}_n^*}(x)$ with $\eta = \epsilon^2 (u_m - u_1)^2/4 \leq 1/n - \epsilon^2/8$. Thus,

$$\prod_{i=1}^n \frac{f_{H_m}(X_i)}{f_{\widehat{G}_n^*}(X_i)} \geq e^{-n\epsilon^2/8}(1-\eta)^n \geq e^{-n(\epsilon^2/8 + \eta)} \geq e^{-1}. \tag{2.115}$$

Let $\mathscr{H}_m$ be the set of all distributions with support $\{u_1, \ldots, u_m\}$ and $\widehat{G}_n = \sum_{j=1}^m \widehat{w}_j^{(k)} \delta_{u_j}$. The upper bound in [18, 71] and (2.87) provide

$$\sup_{H \in \mathscr{H}_m} \prod_{i=1}^n \frac{f_H(X_i)}{f_{\widehat{G}_n}(X_i)} \leq \max_{j \leq m} \Big(\frac{w_j^{(k)}}{w_j^{(k-1)}} \Big)^n \leq \frac{1}{e q_n}.$$

This and (2.115) imply $\prod_{i=1}^n f_{\widehat{G}_n^*}(X_i) \leq q_n^{-1} \prod_{i=1}^n f_{\widehat{G}_n}(X_i)$. $\square$

Chapter 3

General Maximum Likelihood Empirical Bayes Wavelet Method and Exactly Adaptive Minimax Estimation

3.1 Introduction

Consider the nonparametric regression model

$$Y_i = f(t_i) + e_i, \quad i = 1, \ldots, N, \tag{3.1}$$

where $t_i = i/N$, and e_i are iid $N(0, \sigma^2)$. We wish to recover the unknown function f based on the sample $\boldsymbol{Y} \equiv (Y_i, i = 1, \ldots, N)$. For example, how do we recover a piecewise polynomial with unknown number and locations of discontinuities? In general, we would like to consider the estimation of a regression function f with unknown discontinuities or inhomogeneous smoothness across different parts of a domain. Through a discrete wavelet transform (DWT) the nonparametric regression problem can be turned into a problem of estimating the wavelet coefficients at individual resolution levels. The estimation at a single resolution level can be treated by considering a more fundamental problem, that is, compound estimation of a vector of normal means. From many points of view, the normal mean problem occupies the heart of statistical estimation theory. It has been considered as the canonical model or motivating example in the developments of adaptive nonparametric regression, empirical Bayes, admissibility, variable selection, multiple testing and many other areas in statistics.

Nonparametric regression is typically studied under smoothness conditions on the known regression function f. Such smoothness conditions have interpretation as sparsity of wavelet coefficients in the sense of having many (near) zeros. Sparse vectors of wavelet coefficients can be treated as members of Besov balls with a small sparsity index $p > 0$. For sparse means, the linear estimators, e.g., the James-Stein estimator, do not achieve the optimal rates of minimax risk [21, 24]. Many wavelet threshold methods have been proposed and proved to be highly adaptive. These threshold methods include the universal threshold estimator [22], and adaptive procedures SURE [23], FDR [1], and parametric EB posterior median (EBThresh) [47, 48]. Block threshold methods have also been considered, e.g., by Cai [12] and Cai and Zhou [15].

Adaptive threshold estimators can be viewed as approximations of an optimal candidate in certain families of separable threshold functions. Instead of restricting the approximation in a particular function family, general empirical Bayes (EB), a greedier approach proposed earlier by Robbins [59, 60], aims to attain the oracle performance of the optimal rule within the class of all separable estimation functions. Here a separable estimator is one that uses fixed deterministic function to estimate all wavelet coefficients within individual resolution levels. Thus, the general EB is greedier in the sense of aiming at the smaller benchmark risk than adaptive threshold methods. This naturally raises the question that whether the gain by aiming at the smaller general EB benchmark is large enough to offset the additional cost of having to pick from a nonparametric family of estimation functions.

Jiang and Zhang [45] proposed a general maximum likelihood EB (GMLEB) method for compound estimation of normal means. They treat the unknown means as iid variables with a completely unknown common “prior” distribution, estimate this nominal prior with the generalized MLE [49], and then use the Bayes rule for the estimated prior. The results there affirm that by aiming at the minimum risk of all separable estimators, the greedier general EB approach realizes significant risk reduction over state-of-the-art threshold methods for the unknown signal vectors of different degrees of sparsity with moderate and large samples.

In this chapter we develop the GMLEB wavelet method in nonparametric regression. We transform the problem in the function domain into to the sequence domain of estimating the wavelet coefficients by DWT, and then apply the GMLEB estimator to observed wavelet coefficients in individual resolution levels. Both the numerical performance and asymptotic properties of the GMLEB are studied. We provide an oracle inequality, that is, an upper bound for the estimation regret. Moreover, it is shown that the GMLEB simultaneously achieves the exactly adaptive minimaxity over all Besov balls, without prior knowledge of the smoothness index of the underlying function. As mentioned earlier, adaptive minimaxity implies the adaptation to spatial inhomogeneity of the unknown function. We conduct an extensive Monte Carlo simulation study of the performance of our estimator with four standard test functions and two signal-to-noise levels. It turns out that our procedure has superior finite sample performance in comparison to the other leading wavelet threshold estimators and a Fourier EB estimator [74, 76]. Applications to the high-throughput screening (HTS) data are used to explore the practical performance of the approach.

This chapter is organized as follows. In Section 3.2, we present the wavelet transform approach and the Besov constraints over unknown functions. We review the GMLEB estimator and implement it in nonparametric regression models in Section 3.3. We state the main theoretical properties of the GMLEB wavelet estimators in Section 3.4. We investigate the practical performance of the proposed method by simulation in Section 3.5. A real data set is considered in Section 3.6. Section 3.7 contains the mathematical proofs of the main theorems.

3.2 Problem Formulation

We introduce some notations used throughout this chapter. Suppose the sample size is $N = 2^{J+1}$ for some integer $J > 0$. Let $\boldsymbol{f}_N \equiv (f(t_i), i = 1, \ldots, N)$ and $\widehat{\boldsymbol{f}}_N \equiv (\widehat{f}(t_i), i = 1, \ldots, N)$ denote the vectors of true and estimated functions respectively. Let $\|\boldsymbol{x}\|^2 \equiv \sum_{i=1}^N x_i^2$ be the ℓ_2 norm. We measure the performance of $\widehat{\boldsymbol{f}}_N$ under the mean squared error (MSE)

$$R^{(N)}(\widehat{\boldsymbol{f}}_N, \boldsymbol{f}_N) = N^{-1}E\|\widehat{\boldsymbol{f}}_N - \boldsymbol{f}_N\|^2 = N^{-1}E\sum_{i=1}^{N}(\widehat{f}(t_i) - f(t_i))^2 \tag{3.2}$$

for any given estimator $\widehat{\boldsymbol{f}}_N$. Although the notation f suggests a function of a real variable t, in this chapter we work only with the sample points t_i.

Most of wavelet-based approaches to the nonparametric regression estimation of $\boldsymbol{f}_N$ proceed by taking the DWT of the data Y_i, processing the noisy wavelet coefficients to estimate the true discrete wavelet coefficients, and then transforming back to obtain the estimate $\widehat{\boldsymbol{f}}_N$. The underlying notion behind wavelet method is that the unknown function has an economical wavelet expression, that is, the large coefficients occur mostly around the spatial inhomogeneities of the unknown function [26]. Hence, the regression function f can be well approximated by estimating a small proportion of relatively large wavelet coefficients.

3.2.1 Wavelet transform

For any $f \in L_2(\mathbb{R})$, wavelet transform is based on translations and dilations of two basis functions called the scaling function ϕ and the mother wavelet ψ. It can be written as

$$f(t) = \sum_{k=1}^{2^{j_0}} \widetilde{\beta}_{j_0k}\phi_{j_0k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=1}^{2^j} \beta_{jk}\psi_{jk}(t), \tag{3.3}$$

where j indicates the resolution level associated with frequency and k indicates the location. The wavelet coefficients are given by $\widetilde{\beta}_{jk} = \int f(t)\phi_{jk}(t)dt$ and $\beta_{jk} = \int f(t)\psi_{jk}(t)dt$. In (3.3), $\widetilde{\beta}_{j_0k}$ are the coefficients at the coarsest level representing the gross structure of the function f, and β_{jk} are the wavelet coefficients which representing finer structures of the function f as the resolution level j increases.

An orthonormal wavelet basis has an associated exact orthogonal DWT. Suppose $N = 2^{J+1}$, a DWT of $\boldsymbol{Y}$ yields empirical wavelet coefficients vector $\boldsymbol{y}$ via

$$\boldsymbol{y} = N^{-1/2}\mathcal{W}\boldsymbol{Y} \tag{3.4}$$

where $\mathcal{W}$, called the finite wavelet transformation matrix, is a real $N \times N$ orthonormal matrix. Write $\boldsymbol{y} = (\widetilde{y}_{j_01}, \ldots, \widetilde{y}_{j_02^{j_0}}, y_{j_01}, \ldots, y_{j_02^{j_0}}, \ldots, y_{J1}, \ldots, y_{J2^J})$. Here $\widetilde{y}_{j_0k}$ are the observed gross structure terms at the lowest resolution level, and the coefficients y_{jk}, $(j = j_0, \ldots, J, k = 1, \ldots, 2^j)$ are observed fine structure wavelet terms.

The DWT (3.4) transforms the problem in the function domain into a problem of estimating the wavelet coefficients in the sequence domain. Let $\boldsymbol{\beta}_N \equiv N^{-1/2}\mathcal{W}\boldsymbol{f}_N$ be the DWT of unknown $\boldsymbol{f}_N$ and denote

$$\boldsymbol{\beta}_N = (\widetilde{\beta}_{j_01}, \ldots, \widetilde{\beta}_{j_02^{j_0}}, \beta_{j_01}, \ldots, \beta_{j_02^{j_0}}, \ldots, \beta_{J1}, \ldots, \beta_{J2^J}).$$

Here β_{jk} is approximately the true wavelet coefficient $\int f(t)\psi_{jk}(t)dt$ of f. In the wavelet domain, we observe the noisy wavelet coefficients y_{jk} up to level J

$$y_{jk} = \beta_{jk} + z_{jk}\sigma_N, \quad j = j_0, \ldots, J, \quad k = 1, \ldots, 2^j, \tag{3.5}$$

where z_{jk} are independent standard normal random variables and $\sigma_N = \sigma/\sqrt{N}$. We wish to estimate $\boldsymbol{\beta}_N$ with small squared error loss $\|\widehat{\boldsymbol{\beta}}_N - \widehat{\boldsymbol{\beta}}_N\|^2$. Applying the inverse DWT, we obtain the estimate of f at the sample points. That is, $\boldsymbol{f}_N$ is estimated by $\widehat{\boldsymbol{f}}_N = N^{1/2}\mathcal{W}^T\widehat{\boldsymbol{\beta}}_N$. The estimate of the whole function f is given by

$$\widehat{f}_N(t) = \sum_{k=1}^{2^{j_0}} \widehat{\widetilde{\beta}}_{j_0k}\phi_{j_0k}(t) + \sum_{j=j_0}^{J}\sum_{k=1}^{2^j} \widehat{\beta}_{jk}\psi_{jk}(t). \tag{3.6}$$

By the Parseval identity, we have $N^{-1}\|\widehat{\boldsymbol{f}}_N - \boldsymbol{f}_N\|^2 = \|\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N\|^2$.

Nonparametric regression model (3.1) is closely related to the white noise model in which we observe a stochastic process

$$Y(t) \equiv \int_0^t f(u)du + \epsilon W(t), \quad 0 \le t \le 1, \tag{3.7}$$

where $f \in L^2[0,1]$ is unknown and $W(\cdot)$ is a standard Brownian motion. The noise level between the two models matches with $\epsilon^2 = \sigma^2/N$. In the white noise model, a wavelet coefficients sequence $y_{jk} = \int \psi_{jk}dY(t) \sim N(\beta_{jk}, \epsilon^2)$ of infinite length is observed while in nonparametric regression, coefficients are observed only up to level J.

3.2.2 The Besov constraints over the unknown function

Following the classical way to study the adaptivity of wavelet smoothing methods, we shall study the worst behavior when the wavelet coefficients are constrained to lie in a particular Besov ball, corresponding to Besov function space membership of the function itself. We shall show that the GMLEB method automatically achieves the exact minimax risks simultaneously in all Besov balls. Especially, this exactly adaptive minimax in Besov balls with small parameter $p > 0$ allows the spatial adaptation to unknown discontinuities or inhomogeneous smoothness in f.

Besov balls with different parameters allow varying degrees of smoothness in the functions which they contain since wavelet coefficients can measure global smoothness. Roughly speaking, the Besov ball $B^{\alpha}_{p,q}$ contains functions having α bounded derivatives in L^p norm. Full details of Besov balls are given, for example, in [69]. The Besov norm of the wavelet coefficients of a function f is

$$\|\boldsymbol{\beta}\|^{\alpha}_{p,q} \equiv \left\{ \sum_{j=j_0}^{\infty} \left(2^{j(\alpha+1/2-1/p)} \Big(\sum_{k=1}^{2^j} |\beta_{jk}|^p \Big)^{1/p} \right)^q \right\}^{1/q}. \tag{3.8}$$

Note that the Besov function norm of index (α, p, q) of a function f is equivalent to the sequence norm (3.8) of the wavelet coefficients of the function. See [53]. The Besov ball is

$$B^{\alpha}_{p,q}(C) \equiv \left\{ \boldsymbol{\beta}\colon\ \|\boldsymbol{\beta}\|^{\alpha}_{p,q} \le C \right\}. \tag{3.9}$$

Since the sequence $\boldsymbol{f}_N$ is of primary interest, we place the Besov restriction on the discrete wavelet coefficients $\boldsymbol{\beta}_N = N^{-1/2}\mathcal{W}\boldsymbol{f}_N$. The constraint $\boldsymbol{\beta}_N \in B^{\alpha}_{p,q}(C)$ depends on both the function f and N. Our asymptotic minimaxity theorem should be thought of as a "triangular array" result for $\boldsymbol{f}_N$, rather than a limiting result for f. With the notation $\boldsymbol{\beta}_N \in B^{\alpha}_{p,q}(C)$, we automatically treat $\beta_{jk} = 0$ when $j > J$ since $\boldsymbol{\beta}_N$ has only $N = 2^{J+1}$ coordinates. Let $\mathscr{R}(B^{\alpha}_{p,q}(C)) = \inf_{\widehat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)} E \sum_{j=j_0}^{\infty} \sum_{k=1}^{2^j} (\widehat{\beta}_{jk} - \beta_{jk})^2$ be the minimax risk of estimating f over the Besov ball $B^{\alpha}_{p,q}(C)$. Donoho and Johnstone [24] show that $\mathscr{R}(B^{\alpha}_{p,q}(C)) \asymp N^{-\alpha/(\alpha+1/2)}$. Moreover, by Hölder inequality,

$\sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)} \sum_{j>J} \sum_k \beta_{jk}^2 \asymp N^{-2(\alpha+1/2-1/p)}$. Thus, condition $\alpha + 1/2 - 1/p > \alpha/(2\alpha+1)$, that is,

$$\frac{2\alpha^2}{2\alpha+1} > \frac{1}{p} - \frac{1}{2} \tag{3.10}$$

allows the term for $j > J$ negligible with respect to the minimax risk, which is necessary for the discussion of the minimax risk rate with only N samples.

3.2.3 Adaptation to inhomogeneous smoothness of the unknown regression function

The idea of adaptive nonparametric regression aims at recovering regression functions with unknown spatial inhomogeneities. Such adaptation to inhomogeneous smoothness is achieved through adaptation to minimax risks in Besov balls with different smoothness and sparse indices. In this subsection, we formally quantify the notion of adaptation to inhomogeneous smoothness.

Let $\mathscr{F}_{d,m}(C)$ be the collection of all piecewise polynomials f of degree d in $[0,1]$, with at most m pieces and $\|f\|_\infty \le C$. Let ψ be a mother wavelet with $\int \psi(t)dt = 0$ and $\psi(t) = 0$ outside an interval I_0 of length $|I_0|$. For $f \in \mathscr{F}_{d,m}(C)$, the wavelet coefficients $\beta_{jk} = \int f(t)\psi_{jk}(t)dt = 2^{j/2}\int f(t)\psi(2^j t - k)dt = 0$ if ψ_{jk} does not contain any discontinuous point and $|\beta_{jk}| \le 2^{-j/2}C\int|\psi(t)|dt$ otherwise. Thus, $\|\boldsymbol{\beta}_{[j]}\|_p \le 2^{-j/2}m^{1/p}CM_0$ where $M_0 = (|I_0|+1)^{1/p}\int|\psi(t)|dt$. By (3.8), $\|\boldsymbol{\beta}\|^{\alpha}_{p,q} \le \infty$ if $\alpha < 1/p$ for $q < \infty$. Combining $\alpha < 1/p$ with (3.10) leads to $\alpha < 1/p < 2\alpha^2/(2\alpha+1) + 1/2$. This example enlighten us to express the adaptation to inhomogeneous smoothness in definition below.

Definition 3.1. *For nonparametric regression model (3.1), an estimator $\widehat{\boldsymbol{f}}_N$ is adaptive to inhomogeneous smoothness of the unknown regression function f if the exactly adaptive minimaxity*

$$\sup_{\boldsymbol{f}_N\in B} R^{(N)}(\widehat{\boldsymbol{f}}_N, \boldsymbol{f}_N) = (1+o(1)) \inf_{\widehat{\boldsymbol{f}}_N} \sup_{\boldsymbol{f}_N\in B} R^{(N)}(\widehat{\boldsymbol{f}}_N, \boldsymbol{f}_N) \tag{3.11}$$

holds in all Besov balls $B^{\alpha}_{p,q}(C)$ *satisfying*

$$\alpha < \frac{1}{p} < \frac{2\alpha^2}{2\alpha+1} + \frac{1}{2}, \tag{3.12}$$

where $R^{(N)}(\widehat{\boldsymbol{f}}_N, \boldsymbol{f}_N)$ *is as in (3.2).*

3.3 The GMLEB Wavelet Method

As mentioned in Section 3.2.1, through DWT the nonparametric regression problem can be turned into a problem of estimating the wavelet coefficients at individual resolution levels. The function estimation procedure as well as the analysis become clear once the problem of estimating the wavelet coefficients at a given resolution level is well understood. The estimation at a single resolution level can be treated by considering a more fundamental problem, that is, compound estimation of a vector of normal means.

The general maximum likelihood empirical Bayes (GMLEB) method for the compound estimation of normal means is considered in detail by [45]. There, they showed that the GMLEB outperforms the James-Stein and several state-of-the-art threshold estimators in a wide range of settings. In this section, we shall briefly review the basic method presented there and describe how to construct the GMLEB wavelet estimator. We divide the section into 2 subsections to describe (1) the general EB and the GMLEB method and (2) the GMLEB wavelet method.

3.3.1 Empirical Bayes and the GMLEB method

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_n)$ are observations satisfying

$$X_i = \theta_i + z_i, \tag{3.13}$$

where z_i are independent standard normal random variables. Compound estimation of normal means concerns the estimation of the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ under the compound squared loss $L_n(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = n^{-1}\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = n^{-1}\sum_{i=1}^{n}(\widehat{\theta}_i - \theta_i)^2$ for any estimation rule $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)$. The estimator $\widehat{\theta}_i \colon \boldsymbol{X} \to \mathbb{R}$ is separable if $\widehat{\theta}_i$ is a fixed deterministic function only of X_i. The compound estimation of a vector of deterministic normal means is closely related to the Bayes estimation of a single random mean. In the Bayes estimation problem,

$$Y|\lambda \sim N(\lambda, 1), \quad \lambda \sim G, \tag{3.14}$$

and we estimate the parameter λ based on the univariate observation Y. The prior distribution $G = G_n$ here which naturally matches the unknown means $\boldsymbol{\theta}$ is the empirical distribution $G_n(u) = G_{n,\boldsymbol{\theta}}(u) = n^{-1}\sum_{i=1}^n I(\theta_i \leq u)$.

In the context of the squared loss, the fundamental theorem of compound estimation [59] asserts that the compound risk of a separable rule $\widehat{\boldsymbol{\theta}} = t(\boldsymbol{X})$ in the multivariate model (3.13) is identical to the mean squared error of the same rule $\widehat{\lambda} = t(Y)$ under the prior of empirical distribution G_n in the univariate model (3.14):

$$E_{\boldsymbol{\theta}} L_n(t(\boldsymbol{X}), \boldsymbol{\theta}) = \int E_\lambda (t(Y) - \lambda)^2 dG_n(\lambda). \tag{3.15}$$

For any true or nominal prior G, the optimal Bayes rule is

$$t^*_G(Y) = \arg\min_t \int E_\lambda (t(Y) - \lambda)^2 dG(\lambda) = \frac{\int u\varphi(Y-u)G(du)}{\int \varphi(Y-u)G(du)} = Y + \frac{f'_G(Y)}{f_G(Y)}, \tag{3.16}$$

where φ is the standard normal density, $f_G(y) \equiv \int \varphi(y-u)G(du)$ is the density of the normal location mixture by distribution G, and $f'_G(y) \equiv df_G(y)/dy$. The minimum Bayes risk is $R^*(G) = \int E_\lambda(t^*_G(Y) - \lambda)^2 dG(\lambda) = 1 - \int (f'_G/f_G)^2 f_G dy$. It follows from (3.15) that among all separable rule, the compound risk is minimized by the Bayes rule t^*_G in (3.16) with prior $G = G_n$, resulting in the general EB benchmark $R^*(G_n)$. The general EB approach seeks procedures which approximate the Bayes rule $t^*_{G_n}$ or approximately achieve the risk benchmark $R^*(G_n)$.

As a natural approach, we consider using the estimation rule $t(\cdot) = t^*_{\widehat{G}_n}(\cdot)$ with a suitable estimate $\widehat{G}_n$ of G_n based on $\boldsymbol{X}$. The GMLEB method [45] replaces the unknown nominal prior G_n of the oracle rule $t^*_{G_n}$ by its generalized MLE [49]

$$\widehat{G}_n = \arg\max_{G \in \mathscr{G}} \prod_{i=1}^n f_G(X_i) \tag{3.17}$$

where $\mathscr{G}$ is the family of all distributions and

$$f_G(x) = \int \varphi(x-u)G(du) \tag{3.18}$$

is the normal mixture with respect to G. Formally, the GMLEB estimator is defined as

$$\widehat{\theta}_i = t^*_{\widehat{G}_n}(X_i) = X_i + \frac{f'_{\widehat{G}_n}(X_i)}{f_{\widehat{G}_n}(X_i)}, \quad i = 1, \ldots, n, \tag{3.19}$$

where t^*_G is the Bayes rule in (3.16), $\widehat{G}_n$ is the generalized MLE (3.17) and $f_G(x)$ is as in (3.18).

The estimator (3.17) is called the generalized MLE since the likelihood is used only as a vehicle to generate the estimator. Here $G = G_n$ is a nominal prior in that the unknown θ_i are assumed to be deterministic parameters instead of random samples from the nominal prior G_n. Thus, the mixture density f_{G_n} is used for the purpose of deriving the GMLEB instead of being the marginal density of X_i.

3.3.2 The GMLEB wavelet method

With the general EB approach for compound estimation of normal means described in Section 3.3.1, wavelet regression at a single resolution level is the case of estimating a vector of normal means, but with unknown common variance.

For $j \geq j_0$, denote $\boldsymbol{\beta}_{[j]} = (\beta_{jk}, k = 1, \ldots, 2^j)$ and $\boldsymbol{y}_{[j]} = (y_{jk}, k = 1, \ldots, 2^j)$ so that $\boldsymbol{\beta}_N = (\widetilde{\boldsymbol{\beta}}_{[j_0]}, \boldsymbol{\beta}_{[j_0]}, \ldots, \boldsymbol{\beta}_{[J]})$ and $\boldsymbol{y} = (\widetilde{\boldsymbol{y}}_{[j_0]}, \boldsymbol{y}_{[j_0]}, \ldots, \boldsymbol{y}_{[J]})$. As in model (3.5), we consider the estimation of $\boldsymbol{\beta}_{[j]}$ under the compound squared loss based on independent observations $\boldsymbol{y}_{[j]}$. Let $(\theta_{jk}, x_{jk}) \equiv (\beta_{jk}, y_{jk})/\sigma_N$ be the standardized parameters and observations with unit variance and $G_{[j]}(u) = n_j^{-1} \sum_{k=1}^{2^j} I(\theta_{jk} \leq u)$ be the empirical distribution of $\boldsymbol{\theta}_{[j]} = \boldsymbol{\beta}_{[j]}/\sigma_N$. Based on Section 3.3.1, the GMLEB estimator of $\boldsymbol{\beta}_{[j]}$ is $\widehat{\boldsymbol{\beta}}_{[j]} \equiv (\widehat{\beta}_{jk}, k = 1, \ldots, 2^j)$ with the coordinates

$$\widehat{\beta}_{jk} \equiv \widehat{\beta}_{jk}(\boldsymbol{y}_{[j]}) \equiv \sigma_N t^*_{\widehat{G}_{[j]}}(y_{jk}/\sigma_N) = \sigma_N t^*_{\widehat{G}_{[j]}}(x_{jk}). \tag{3.20}$$

where $\widehat{G}_{[j]} = \arg\max_{G \in \mathscr{G}} \prod_{k=1}^{2^j} f_G(x_{jk})$ is the generalized MLE of $G_{[j]}$ as in (3.17), and $t^*_{\widehat{G}_{[j]}}$ is the estimate of $t^*_{G_{[j]}}$.

In (3.16) we may need to avoid dividing by a near zero quantity, which will result in dramatic change in ratio. This notion leads to the following regularized Bayes estimator

$$t^*_G(x;\rho) = x + \frac{f'_G(x)}{f_G(x) \vee \rho}. \tag{3.21}$$

For $\rho = 0$, $t^*_G(x;0) = t^*_G(x)$ is the Bayes estimator for the prior G. For $\rho = \infty$, $t^*_G(x;\infty) = x$ gives the classical MLE which requires no knowledge of the prior. Let $n_j = 2^j$. As stated in Proposition 2 of [45], the connection between the GMLEB estimator (3.19) and the regularized Bayes rule (3.21) is provided by

$$t^*_{\widehat{G}_{[j]}}(x_{jk}) = t^*_{\widehat{G}_{[j]}}(x_{jk};\rho_{n_j}), \quad \rho_{n_j} = q_{n_j}/(en_j\sqrt{2\pi}), \quad q_{n_j} = (e\sqrt{2\pi}/n_j^2) \wedge 1, \tag{3.22}$$

when the approximated generalized MLE $\widehat{G}_{[j]}$ satisfies that

$$\prod_{k=1}^{n_j} f_{\widehat{G}_{[j]}}(x_{jk}) \geq q_{n_j} \sup_{G\in\mathscr{G}} \prod_{k=1}^{n_j} f_G(x_{jk}), \quad \widehat{G}_{[j]} \in \mathscr{G}, \tag{3.23}$$

where $\mathscr{G}$ is the family of all distribution functions and $x_{jk} = y_{jk}/\sigma_N$. Thus, when condition (3.23) holds, the regularized GMLEB (3.21) is identical to the GMLEB (3.20). The purpose to represent the GMLEB estimator as a regularized one is to facilitate the theoretical investigation so that an oracle inequality which provides a uniform upper bound of the regret is derived, see [45].

In view of (3.20) and (3.21), we construct a GMLEB wavelet estimator in the multi-resolution analysis problem (3.5) for $j \geq j_0$.

$$\widehat{\boldsymbol{\beta}}_{[j]} \equiv \widehat{\boldsymbol{\beta}}(\boldsymbol{y}_{[j]}) \equiv \{\widehat{\beta}_{jk}\}, \quad \widehat{\beta}_{jk} \equiv \sigma_N t^*_{\widehat{G}_{[j]}}(y_{jk}/\sigma_N;\rho_{n_j}), \tag{3.24}$$

where $t^*_G(\cdot;\rho)$ is as in (3.21), and ρ_{n_j} is as in (3.22). For unknown σ_N, estimator of σ_N can be constructed from the median absolute deviations (MAD) of the observations at the highest resolution level, that is,

$$\widehat{\sigma}_N \equiv \text{MAD}(\boldsymbol{y}_{[J]}) \equiv \frac{\text{median}(|y_{Jk}|: 1 \leq k \leq 2^J)}{\text{median}(|N(0,1)|)}. \tag{3.25}$$

We estimate the coefficients β_{jk} for $j \geq j_0$ level by level by the estimate in (3.24). The coefficients $\widetilde{\beta}_{j_0k}$ are estimated by their observed values $\widetilde{y}_{j_0k}$. So

$$\widehat{\boldsymbol{\beta}}_N = (\widetilde{y}_{j_01}, \ldots, \widetilde{y}_{j_02^{j_0}}, \widehat{\beta}_{j_01}, \ldots, \widehat{\beta}_{j_02^{j_0}}, \ldots, \widehat{\beta}_{J1}, \ldots, \widehat{\beta}_{J2^J}). \tag{3.26}$$

To obtain the estimates $\widehat{f}_N(t_i)$ of the function values $f(t_i)$, apply the inverse DWT, $\widehat{\boldsymbol{f}}_N = \sqrt{N}\mathcal{W}^T\widehat{\boldsymbol{\beta}}_N$.

3.4 The Oracle Inequality and Its Consequences

In this section we state the concepts of uniform ideal adaptivity and exactly adaptive minimaxity. Also, we shall derive these properties of the GMLEB wavelet method.

3.4.1 An oracle inequality

Consider the estimation of unknown wavelet coefficients $\boldsymbol{\beta}_N$ based on observed wavelet coefficients $\boldsymbol{y}$ following that $y_{jk} \sim N(\beta_{jk}, \sigma^2/N)$, $j = j_0, \ldots, J$, $k = 1, \ldots, 2^j$ with $\sigma_N^2 = \sigma^2/N$. For a level-by-level estimator $\widehat{\boldsymbol{\beta}}_N$, denote

$$R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]}, \boldsymbol{\beta}_{[j]}) \equiv E_{\boldsymbol{\beta}_{[j]}} \left\| \widehat{\boldsymbol{\beta}}_{[j]} - \boldsymbol{\beta}_{[j]} \right\|^2, \tag{3.27}$$

$$R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) \equiv \sum_{j=j_0}^{J} R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]}, \boldsymbol{\beta}_{[j]}), \tag{3.28}$$

so that $R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]}, \boldsymbol{\beta}_{[j]})$ and $R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N)$ are the ℓ_2 risk for the j-th resolution level and the total levels, respectively. An oracle expert with the knowledge of $t^*_{G_{[j]}}$ could use the ideal separable rule $\sigma_N t^*_{G_{[j]}}(y_{jk}/\sigma_N)$ in (3.20) for β_{jk} to achieve the ideal risk

$$R^{(N,*)}(\boldsymbol{\beta}_{[j]}) \equiv \min_{\widehat{\boldsymbol{\beta}}_{[j]} \in \mathscr{D}^s} R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]}, \boldsymbol{\beta}_{[j]}) \equiv \min_{t(\cdot)} E_{\boldsymbol{\beta}_{[j]}} \left\| \sigma_N t(\boldsymbol{y}_{[j]}/\sigma_N) - \boldsymbol{\beta}_{[j]} \right\|^2 \tag{3.29}$$

$$R^{(N,*)}(\boldsymbol{\beta}_N) \equiv \sum_{j=j_0}^{J} \min_{\widehat{\boldsymbol{\beta}}_{[j]} \in \mathscr{D}^s} R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]}, \boldsymbol{\beta}_{[j]}) \equiv \sum_{j=j_0}^{J} R^{(N,*)}(\boldsymbol{\beta}_{[j]}), \tag{3.30}$$

where $\mathscr{D}^s$ is the collection of all separable estimates of the form $\widehat{\beta}_{jk} = t_j(y_{jk})$. Although $\sigma_N t^*_{G_{[j]}}(y_{jk}/\sigma_N)$ are not statistics, the ideal risk (3.30) provides a benchmark for each level in our problem. Theorem 3.1 provides a crucial oracle inequality in the derivation of our main results. It allow us to bound the maximum regret of our estimator in all Besov balls.

Theorem 3.1. *Let $\widehat{\boldsymbol{\beta}}_N$ be the GMLEB estimator (3.26) with approximate generalized MLEs $\widehat{G}_{[j]}$ satisfying (3.23) for all $j_0 \le j \le J$. Let $R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N)$ be the total ℓ_2 risk as in (3.28) and $R^{(N,*)}(\boldsymbol{\beta}_N)$ be the ideal risk in (3.30). Let $B^{\alpha}_{p,q}(C)$ be the Besov ball as in (3.9). Denote $n_j = 2^j$, $j_1 \equiv \max(\inf\{j\colon \log n_j \ge 2/p\}, j_0)$, and ξ_0 as an arbitrary positive constant. Under condition (3.10), there exists a universal constant M such that*

$$\begin{aligned} &\sup_{\boldsymbol{\beta}_N \in B^{\alpha}_{p,q}(C)} \Big\{ R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) - R^{(N,*)}(\boldsymbol{\beta}_N) \Big\} \qquad (3.31) \\ \le\ & M\sigma^2/N \Big\{ \sum_{j=j_0}^{j_1-1} n_j \log n_j + \Big(1 + \frac{1}{\xi_0}\Big) \sum_{j=j_1}^{J} n_j r_p\Big(n_j, \frac{C n_j^{-(\alpha+1/2)}}{\sigma/\sqrt{N}}\Big)\Big\} \\ & + M\xi_0 N^{-\alpha/(\alpha+1/2)} C^{1/(\alpha+1/2)}. \end{aligned}$$

where $r_p(n, D) = (\log n)^5/n + (\log n)^{4+p/(2+2p)}(D/n)^{p/(1+p)}$.

3.4.2 Uniform ideal adaptation

For any class B, the minimax risk for the total squared loss (3.28) is

$$\mathscr{R}^{(N)}(B) \equiv \inf_{\widehat{\boldsymbol{\beta}}_N} \sup_{\boldsymbol{\beta}_N \in B} R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N). \qquad (3.32)$$

We call $\widehat{\boldsymbol{\beta}}_N$ uniformly adaptive to the ideal risk $R^{(N,*)}(\boldsymbol{\beta}_N)$ as in (3.30), with respect to a collection $\mathscr{B}$, if for all $B \in \mathscr{B}$, $\sup_{\boldsymbol{\beta}_N \in B} \Big\{ R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) - R^{(N,*)}(\boldsymbol{\beta}_N)\Big\} = o(1)\mathscr{R}^{(N)}(B)$ where $\mathscr{R}^{(N)}(B)$ is the minimax risk in (3.32). In other words, uniform ideal adaptation demands that, for all $B \in \mathscr{B}$ and in the minimax sense, the regret $r^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) \equiv R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) - R^{(N,*)}(\boldsymbol{\beta}_N)$ to be uniformly of smaller order than the minimax rates in B. The following theorem states that the GMLEB wavelet estimator (3.26) possesses the uniform ideal adaptivity property with respect to all Besov balls as in (3.9).

Theorem 3.2. *Let $\widehat{\boldsymbol{\beta}}_N$ be the GMLEB estimator (3.26) with approximate generalized MLEs $\widehat{G}_{[j]}$ satisfying (3.23) for all $j_0 \le j \le J$. Let $R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N)$ be the total squared risk as in (3.28) and $R^{(N,*)}(\boldsymbol{\beta}_N)$ be the ideal risk in (3.30). Under*

(3.10),

$$\sup_{\boldsymbol{\beta}_N \in B^{\alpha}_{p,q}(C)} \Big\{ R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) - R^{(N,*)}(\boldsymbol{\beta}_N) \Big\} = o(1)\mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)). \tag{3.33}$$

3.4.3 Adaptive minimaxity

A main consequence of the uniform ideal adaptivity is the exactly adaptive minimaxity over all Besov balls. Minimaxity is commonly used to measure the performance of statistical procedures. An estimator is minimax in a specific class B of unknown mean vectors if it attains $\mathscr{R}^{(N)}(B)$, but this does not guarantee satisfactory performance since the minimax estimator is typically uniquely tuned to the specific set B. For small B, the minimax estimator has high risk outside B. For large B, the minimax estimator is too conservative by focusing on the worst case scenario within B. Adaptive minimaxity overcomes this difficulty by requiring $\sup_{\boldsymbol{\beta}_N \in B} R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) = (1+o(1))\mathscr{R}^{(N)}(B)$ simultaneously for all B in certain class $\mathscr{B}$. The adaptive minimaxity in Besov balls with small index $p > 0$ is used to measure the performance of estimators for spatially inhomogeneous function f. The following theorem establishes the exactly adaptive minimaxity of the GMLEB wavelet estimator (3.26).

Theorem 3.3. *Let $\widehat{\boldsymbol{\beta}}_N$ be the GMLEB estimator (3.26) with approximate generalized MLEs $\widehat{G}_{[j]}$ satisfying (3.23) for all $j_0 \le j \le J$. Let $R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N)$ be the total ℓ_2 risk as in (3.28) and $R^{(N,*)}(\boldsymbol{\beta}_N)$ be the ideal risk in (3.30). Under the constraint (3.10), the adaptive minimaxity*

$$\sup_{\boldsymbol{\beta}_N \in B^{\alpha}_{p,q}(C)} R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) = (1+o(1))\mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)) \tag{3.34}$$

holds for all Besov balls.

We translate the exactly adaptive minimaxity (3.34) to the function space. The following theorem is immediate since $N^{-1}\|\widehat{\boldsymbol{f}}_N - \boldsymbol{f}_N\|^2 = \|\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_N\|^2 = 2^{j_0}\sigma^2/N + \sum_{j=j_0}^{J} \|\widehat{\boldsymbol{\beta}}_{[j]} - \boldsymbol{\beta}_{[j]}\|^2$.

Theorem 3.4. *Let $\widehat{\boldsymbol{f}}_N = \sqrt{N}\mathcal{W}^T\widehat{\boldsymbol{\beta}}_N$ be the estimates of $\boldsymbol{f}_N$ based on $\widehat{\boldsymbol{\beta}}_N$ as in (3.26). Under the constraint (3.10),*

$$\sup_{\boldsymbol{f}_N \in B^{\alpha}_{p,q}(C)} R^{(N)}(\widehat{\boldsymbol{f}}_N, \boldsymbol{f}_N) = (1+o(1)) \inf_{\widehat{\boldsymbol{f}}_N} \sup_{\boldsymbol{f}_N \in B^{\alpha}_{p,q}(C)} R^{(N)}(\widehat{\boldsymbol{f}}_N, \boldsymbol{f}_N), \qquad (3.35)$$

where $R^{(N)}(\widehat{\boldsymbol{f}}_N, \boldsymbol{f}_N) = N^{-1}E\|\widehat{\boldsymbol{f}}_N - \boldsymbol{f}_N\|^2$ is as in (3.2). Thus, by Definition 3.1, the GMLEB wavelet estimator (3.26) is adaptive to inhomogeneous smoothness of the unknown function.

3.5 Some Simulation Results

A simulation study is carried out for the nonparametric regression models which are standard in the consideration of wavelet methods. We compare the numerical performance of the GMLEB with that of SURE [23], FDR [1], and EBThresh [47]. SURE is a soft threshold procedure which selects the threshold level at each resolution level by minimizing Stein's unbiased risk estimate. EBThresh is a threshold method based on the posterior median for Gaussian errors with respect to a prior as the mixture of the point mass at zero and a given symmetric distribution. For further details see the original paper.

Four standard test functions, representing different degrees of spatial variability, and various signal-to-noise ratios (SNR) are used for comparison. Sample sizes of $N = 2048$ and $N = 4096$ and SNR of 3 and 7 are considered. The SNR is the ratio of the standard deviation of the function values to the standard deviation of the noise. Johnstone and Silverman [48] reported results of an extensive simulation study of fourteen estimators. In Table 1, we display our simulation results under the same settings as in [48]. Fifteen estimators of various wavelet methods are compared: the James-Stein, the EBThresh using the Laplace posterior median and mean, Cauchy posterior median, the SURE applied to the 4 and 6 highest levels of coefficients, the soft threshold at the universal threshold level $\sqrt{2\log n}$, the FDR at levels $q = 0.01$, 0.05, 0.1 and 0.4, the GMLEB with the uniform initialization, the S-GMLEB with the initialization as in (2.20), the

Table 3.1: Average total squared errors $\widehat{\sigma}^{-2}\|\widehat{\boldsymbol{f}}_N - \boldsymbol{f}_N\|^2$ for $N = 2048$ points for various models and methods. Each entry is based on 100 replications. In each replication, the signal $f(t_i)$ is generated by repeating the original signal function with length 512 four times.

	High noise (SNR = 3)				Low noise (SNR = 7)			
method	bumps	blocks	doppler	heavisine	bumps	blocks	doppler	heavisine
James-Stein	1166	766	644	**142**	1453	1280	1101	320
Laplace (median)	749	616	424	146	753	709	555	250
Cauchy (median)	752	676	425	159	719	657	539	270
Laplace (mean)	**685**	**576**	**387**	**143**	**691**	**640**	**506**	**238**
SURE (4 levels)	970	832	527	185	841	762	828	368
SURE	975	912	514	151	971	955	822	417
Universal soft	3039	1884	1080	266	4554	3065	1917	582
FDR ($q = 0.01$)	1053	889	486	222	906	859	695	335
FDR ($q = 0.05$)	899	758	466	192	808	783	605	290
FDR ($q = 0.1$)	867	726	472	184	807	768	599	282
FDR ($q = 0.4$)	979	810	598	222	1008	939	779	349
GMLEB	**651**	**569**	**371**	150	**642**	**591**	**464**	**243**
S-GMLEB	**648**	**560**	**365**	**144**	**640**	**586**	**461**	**235**
F-GEB	865	772	560	366	857	795	660	443
HF-GEB	744	646	429	149	746	690	558	265

F-GEB and HF-GEB as the Fourier general EB [74] and a hybrid of its monotone version with the EBThresh. Except the SURE applied to the 4 highest resolution levels, all the other methods are applied to the 6 highest resolution levels. When different approaches are used in the wavelet context, the methods are applied separately at each level. In Table 3.1, for each model and noise level, 100 replications are generated. In each replication, the function is generated by repeating 4 times of 512 equally spaced points t_i. The same standard normal noise variables are simulated for each of the models and noise levels in every replication. The error reported here are $\widehat{\sigma}^{-2}\|\widehat{\boldsymbol{f}}_N - \boldsymbol{f}_N\|^2$ where in each realization, the estimated noise variance $\widehat{\sigma}$ is the median absolute deviations of the wavelet coefficients at the highest resolution level. In each column, boldface entries denote the top three estimators other than the hybrid estimator.

Table 3.2: Average total squared errors $\widehat{\sigma}^{-2}\|\widehat{\boldsymbol{f}}_N - \boldsymbol{f}_N\|^2$ for $N = 4096$ points for various models and methods. Each entry is based on 100 replications. In each replication, the signal $f(t_i)$ is generated by repeating the original signal function with length 1024 four times.

	High noise (SNR = 3)				Low noise (SNR = 7)			
method	bumps	blocks	doppler	heavisine	bumps	blocks	doppler	heavisine
James-Stein	1845	1176	862	199	2668	2252	1518	460
Laplace (median)	1118	856	558	**193**	1273	1083	754	**306**
Cauchy (median)	1118	896	572	206	1225	1022	738	307
Laplace (mean)	**1027**	**791**	**511**	**187**	**1149**	**966**	**695**	**290**
SURE (4 levels)	1279	1036	778	325	1402	1497	898	502
SURE	1269	1011	709	220	1399	1486	851	430
Universal soft	4186	2361	1264	220	6268	3987	2344	536
FDR ($q = 0.01$)	1462	1118	650	220	1500	1295	892	359
FDR ($q = 0.05$)	1270	996	626	215	1373	1186	818	335
FDR ($q = 0.1$)	1235	969	633	216	1374	1177	822	335
FDR ($q = 0.4$)	1491	1148	804	292	1750	1473	1086	441
GMLEB	**1033**	**791**	**500**	213	**1115**	**927**	**666**	**306**
S-GMLEB	**1018**	**776**	**483**	**197**	**1104**	**916**	**653**	**291**
F-GEB	1309	1058	776	496	1405	1192	954	580
HF-GEB	1127	867	572	199	1244	1064	770	327

These simulation results can be summarized as follows. The average ℓ_2 loss of the S-GMLEB happens to be the smallest among the fifteen estimators. The S-GMLEB and GMLEB clearly outperform all other methods by large margin except for heavisine. For high noise signals with SNR = 3, the EBThresh with Laplace mean, the S-GMLEB and GMLEB estimators yield comparable results, and they all outperform the Fourier general EB and James-Stein estimators. For the HeaviSine signal, the EBThresh with Laplace mean yields very strong results as competitive as the S-GMLEB. Since the oracle prior (2.20) has a point mass at 0 in all the models used to generate data in this simulation experiment, the S-GMLEB yields slightly better results than the GMLEB as expected.

In Table 3.2 we report simulation results for $n = 4096$. In each replication, the function is generated by repeating 4 times of 1024 equally spaced points

t_i. Compared with Table 3.1, the EBThresh with Laplace median replaces the James-Stein as a third top performer for heavisine function with SNR = 3. The simulations presented here demonstrate the computational feasibility of the proposed GMLEB wavelet method. The strong performance of the both versions of the GMLEB is impressive, since the GMLEB is not specially designed to recover spatial inhomogeneous signals as threshold estimators are.

3.6 Illustrative Data Example

3.6.1 The HTS data

High-throughput screening (HTS) is a large-scale manufacturing process that screens hundreds of thousands to millions of compounds in order to identify potentially leading candidates rapidly and accurately. In HTS, the input is samples to be measured and "reagents" (possibly including membranes, whole cells, or other biological entities as well as chemicals) with which to measure them, and the output is numbers. We show two examples of the HTS data in Figure 3.1. Since the scanning machine measures the difference of certain disease-indicating index, the data points with large negative values indicate the potential leading candidate.

As with any manufacturing process, the output varies. Some of the variability in the results is due to systematic variation in the measurement process. In the bottom panel in Figure 3.1, there is a piece of data located at the down side of the sequence. This may caused by the failure of some experiment devices. This piece of data cannot be considered as the further candidates instead of other scattered outliers, although they are with large absolute negative values. Meanwhile, in Figure 3.1, we can see some baseline curve pattern in each sequence caused by position effect. The baseline pattern will change from sequence to sequence. Our objective is to remove the downside piece of data and the baseline curves.

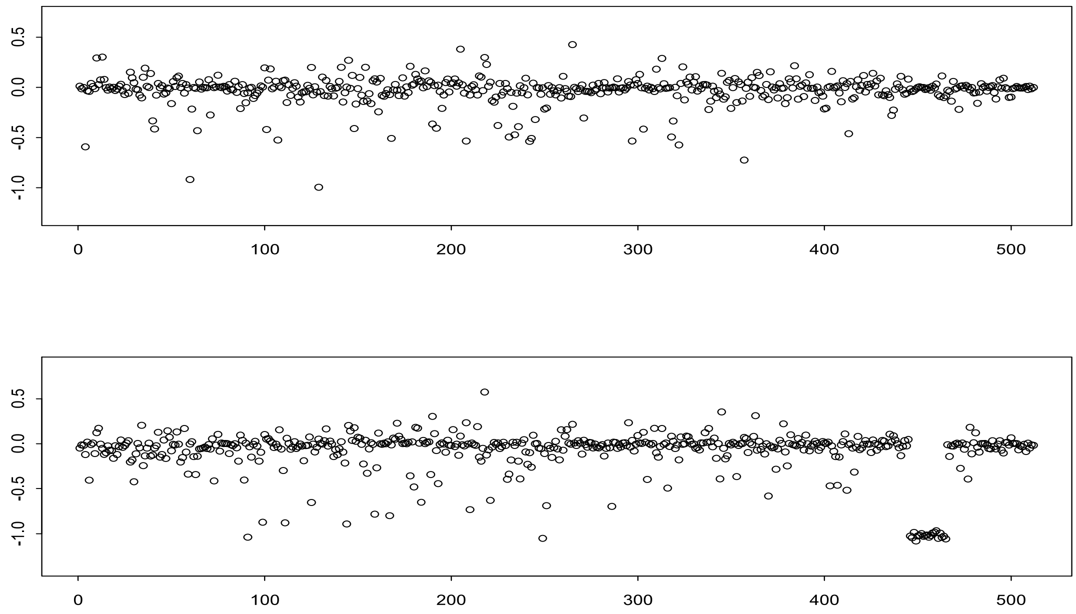

Figure 3.1: Two examples of the HTS data. Each panel corresponds to a particular location among all the plates. Top: a "good" example; bottom: a "bad" example.

3.6.2 Analysis of the HTS data

By the preceding description, for each sequence, we model the data as

$$Y_i = g(t_i) + \mu_i + \sigma z_i, \quad i = 1, \ldots, N, \tag{3.36}$$

where Y_i represents the observed data, the function g represents the baseline curve and, if necessary, the downside piece of data as shown in the bottom panel of Figure 3.1, μ_i represents the true value of the disease-indicating index, and z_i are independent standard normal variables. The data points with large absolute negative value of μ_i are strong candidates. Our objective is to estimate g so that further analysis could be based on the residuals $X_i \equiv Y_i - \widehat{g}(t_i)$.

The GMLEB smoothing technique proceeds as follows. Suppose that $\boldsymbol{y} = N^{-1/2}\mathcal{W}\boldsymbol{Y}$ are the discrete wavelet coefficients of the original sequence $\boldsymbol{Y}$. The coefficients y_{jk} follow the model

$$y_{jk} = \beta_{jk} + z_{jk}\sigma/\sqrt{N}, \tag{3.37}$$

where z_{jk} are independent standard normal random variables. We obtain the

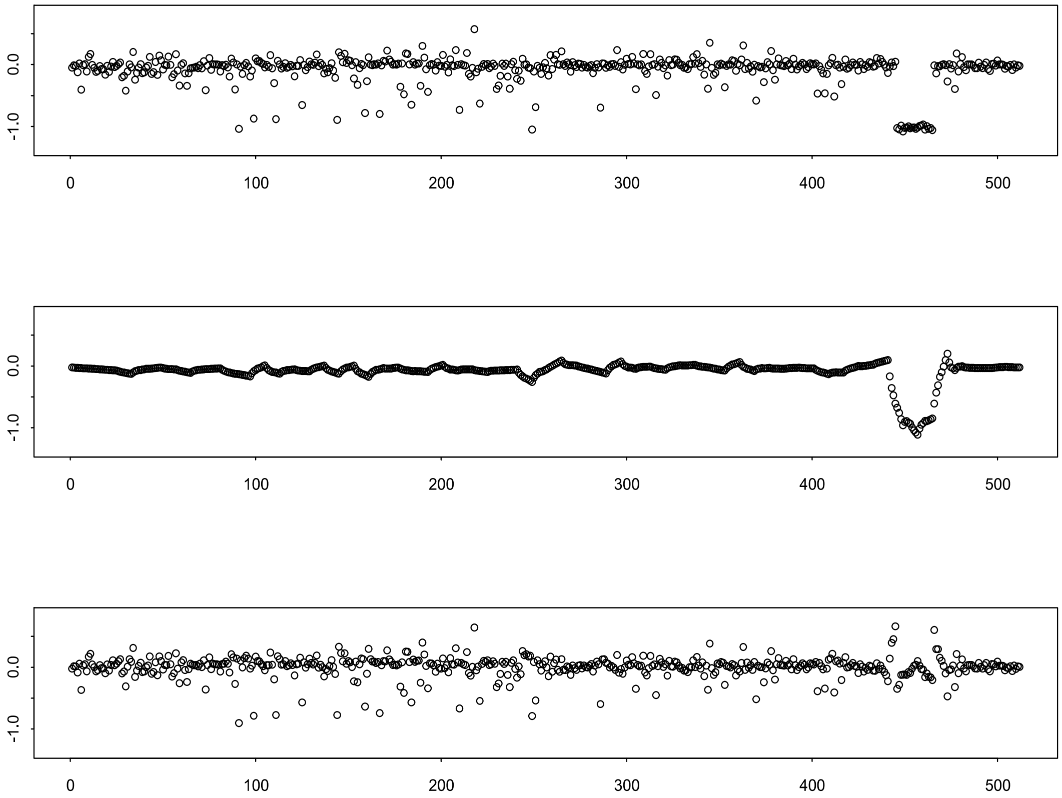

Figure 3.2: The analysis on the bottom sequence in Figure 3.1 by the GMLEB smoothing procedure (3.38), $j_0 = 3$ and $j^* = 5$. Top: the original sequence; middle: the separated baseline pattern and downside piece $\widehat{g}$; bottom: the residuals by subtracting $\widehat{g}$ from the original signal.

estimation $\widehat{g}$ based on the coefficients

$$(\widetilde{y}_{j_0 1}, \ldots, \widetilde{y}_{j_0 2^{j_0}}, \widehat{\beta}_{j_0 1}, \ldots, \widehat{\beta}_{j_0 2^{j_0}}, \ldots, \widehat{\beta}_{j^* 1}, \ldots, \widehat{\beta}_{j^* 2^{j^*}}, 0, \ldots, 0). \tag{3.38}$$

where $\widehat{\beta}_{jk}$ are the estimations by implementing the GMLEB procedure (3.20) level by level for $j_0 \leq j \leq j^*$. Figure 3.2 shows the results by applying the smoothing procedure to the bottom sequence in Figure 3.1. In this example we use the Daubechies' *d4* wavelet basis, with $N = 512$, $j_0 = 3$ and $j^* = 5$.

In Figure 3.2, strong edge effect at the two ends of downside piece of data can be observed. The edge effect means that instead of mimicking the jump points, the bad part is connected with other pieces of sequence in both ends. The reason of the edge effect is that in procedure (3.38), we "kill" all the coefficients in the resolution levels higher than j^*. Thus, we not only remove the random error, but also throw away the true coefficient. Since the coefficients at high levels capture the local feature of a function, ignoring them will result in the loss of the "local information". Since the outliers and the jump points are local features, both of them disappear. Indeed, our purpose is to retain the jump points in $\widehat{g}$ and exclude the outliers.

3.6.3 Removing the edge effect

In this subsection we discuss how to remove the edge effect. Our strategy is to expand the data sequence with respect to some wavelet basis, plot the coefficients at several high resolution levels. There will be large coefficients around the outliers and the discontinuous points. If these two types of large coefficients are different, and moreover, it is possible to design algorithm to classify these two types, then it is promising to remove the edge effect.

To investigate the feasibility of our plan, we first plot some high resolution coefficients. We still work on the bottom sequence in Figure 3.1. In Figure 3.3, we plot coefficients of the five highest resolution levels. As we can see, in the top three levels, large coefficients appear around both outliers and discontinuous points. However, in lower levels, large coefficients only appear around the discontinuous points gradually.

We propose an algorithm below.

1. Set the candidate set C as empty, $C = \varnothing$.

2. Compute the discrete wavelet coefficients by (3.4).

3. For the J-th resolution level, denote $K \equiv \{k\colon |y_{J,k}| \geq \widehat{\sigma}_J z(\alpha_1/2)\}$ where $\widehat{\sigma}_J = \text{MAD}(\boldsymbol{y}_{[J]})$ and z is the right Gaussian quantile. We denote the member of set K as $k_1, \dots, k_m$ where $k_1 < \cdots < k_m$.

4. Set $i = 1$.

5. If (i) $|y_{J,k_1^*}| < \widehat{\sigma}_J z(\alpha_2/2)$ for $k_1^* = k_i + 1, \dots, k_{i+1} - 1$ and (ii) there exists some k_2^*, $\lceil k_i \rceil/8 \leq k_2^* \leq \lceil k_{i+1} \rceil/8$ such that $|y_{J-3,k_2^*}| \geq \widehat{\sigma}_{J-3} z(\alpha_2/2)$, then update $C \leftarrow C \cup \{2k_i, 2k_i + 1, \dots, 2k_{i+1}\}$. Otherwise keep C.

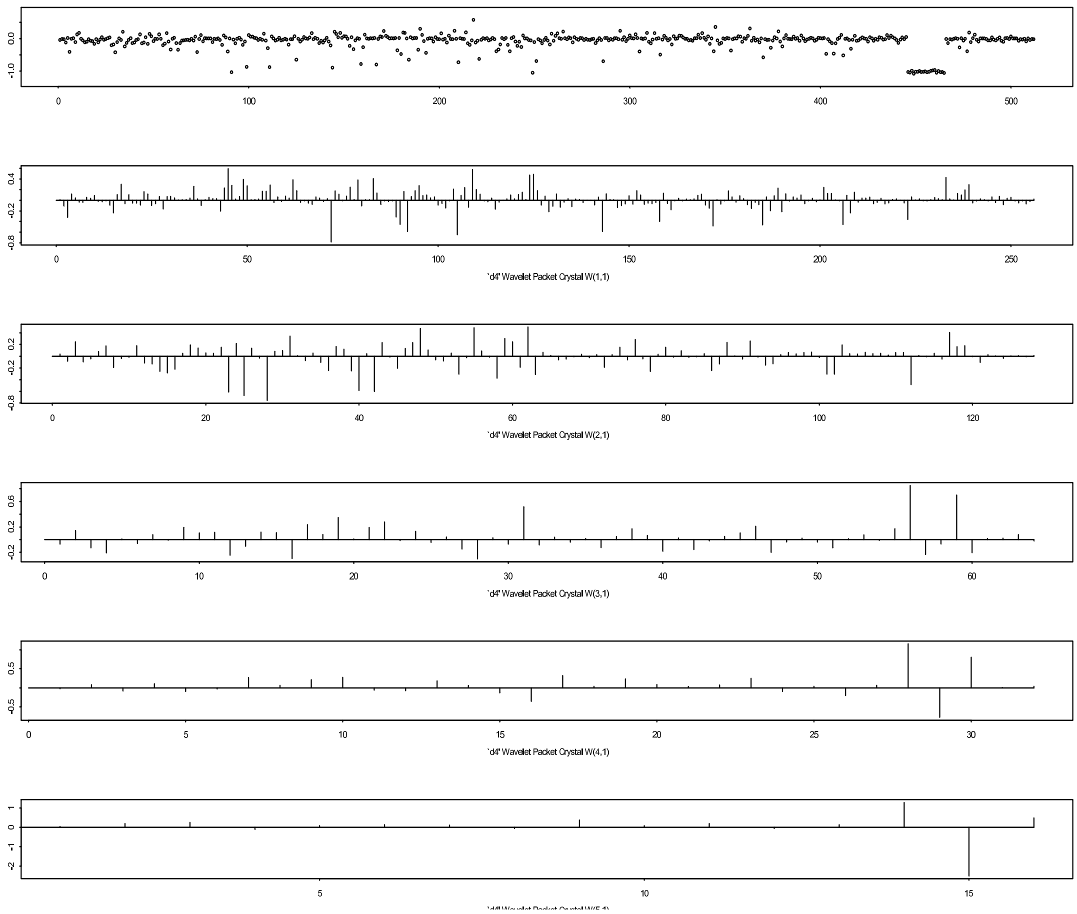

Figure 3.3: The wavelet coefficients of five highest resolution levels of the bottom sequence in Figure 3.1 using the Daubechies' *d4* wavelet basis.

6. Update $i \leftarrow i + 1$ and repeat step 5 until $i = m$.

7. We obtain the estimation $\widehat{g}$ based on the coefficients

$$\begin{cases} \widehat{\beta}_{jk}, & \text{if } j_0 \leq j \leq j^* \text{ or } j > j^* \text{ and } k \in C_j, \\ 0, & \text{if } j > j^* \text{ and } k \notin C_j, \end{cases} \tag{3.39}$$

where $\widehat{\beta}_{jk}$ are the estimations by implementing the GMLEB procedure (3.20) level by level. and $C_j = \lceil C/2^{J-j} \rceil$ where J is the highest resolution level.

There are three tuning parameters in the algorithm. Parameters α_1 and α_2 select large wavelet coefficients by setting a threshold level. Parameter α_3 filters

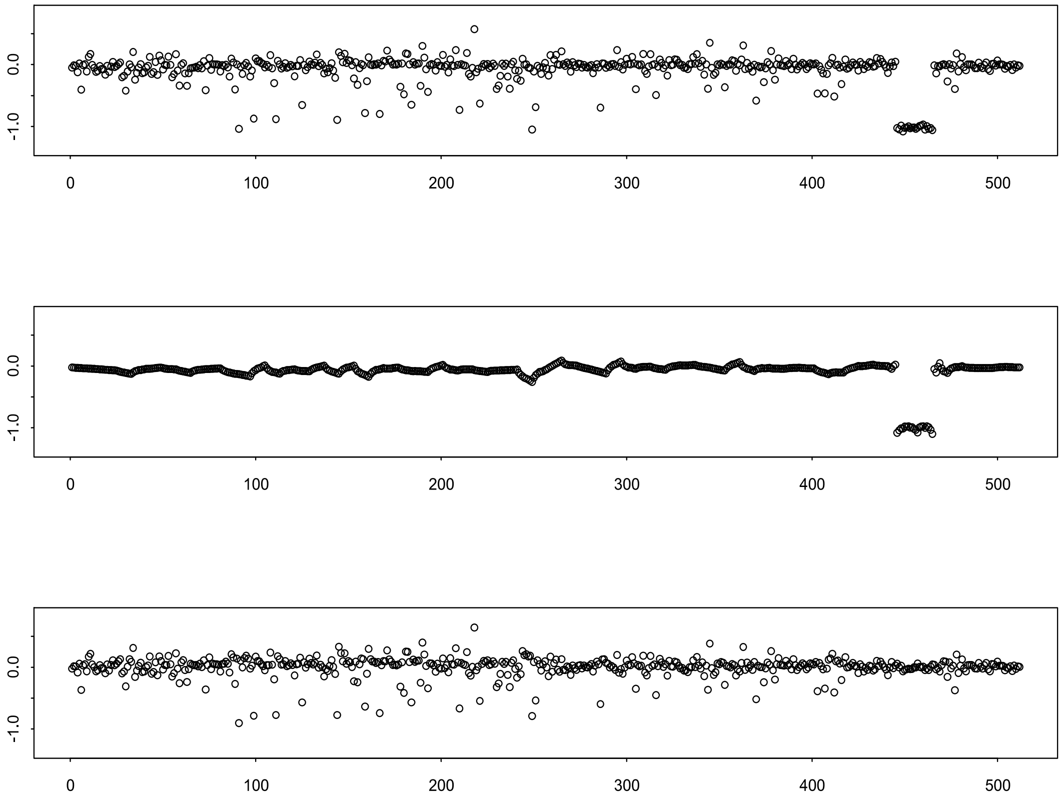

Figure 3.4: Applying the smoothing algorithm to the bottom sequence in Figure 3.1 to remove the edge effect. Top: the original sequence; middle: the separated baseline pattern and downside piece $\widehat{g}$ computed by the proposed smoothing algorithm; bottom: the residuals by subtracting $\widehat{g}$ from the original signal. The parameters are $\alpha_1 = 0.1$, $\alpha_2 = 0.3$, $\alpha_3 = 0.05$, $j_0 = 3$ and $j^* = 5$.

some selected large coefficients by requiring that each coefficient between two neighboring large coefficients to be under certain threshold level.

To explore how the proposed smoothing algorithm works, we conduct some analysis on the bottom sequence in Figure 3.1. The result in Figure 3.4 is very encouraging. With the choices of $\alpha_1 = 0.1$, $\alpha_2 = 0.3$ and $\alpha_3 = 0.05$, the edge effects are successfully removed, compared with the results in Figure 3.2. By removing the edge effect, we avoid introducing new outliers which are caused by the continuous edge.

The algorithm can be generalized to remove multiple edge effects directly. We provide such an example in Figure 3.5. We add artificial errors with different lengths to the same sequence. The algorithm works well since it removes all six

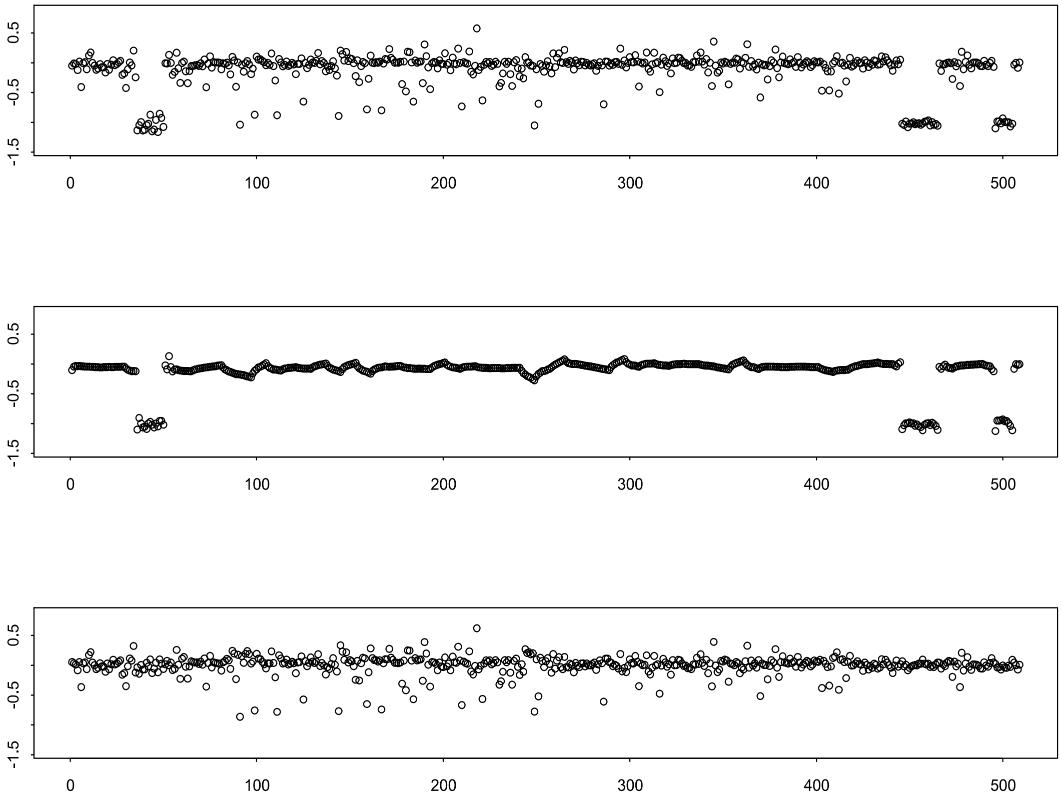

Figure 3.5: Applying the smoothing algorithm to remove the multiple edge effects. Top: the original sequence; middle: the separated baseline pattern and downside piece $\widehat{g}$ computed by the proposed smoothing algorithm; bottom: the residuals by subtracting $\widehat{g}$ from the original signal. The parameters are $\alpha_1 = 0.002$, $\alpha_2 = 0.05$, $\alpha_3 = 0.05$, $j_0 = 3$ and $j^* = 5$.

edges simultaneously. Our parameters are $\alpha_1 = 0.002$, $\alpha_2 = 0.05$, $\alpha_3 = 0.05$, $j_0 = 3$ and $j^* = 5$.

3.7 Proof

We shall use M to denote a universal constant which may take different values from one appearance to another, that is, $M \equiv O(1)$ uniformly.

Proof of Theorem 3.1. For convenience, we denote $A_j = R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]}, \boldsymbol{\beta}_{[j]}) - R^{(N,*)}(\boldsymbol{\beta}_{[j]})$ as the regret of $\widehat{\boldsymbol{\beta}}$ at the j-th level. Let $Y|\lambda \sim N(\lambda, 1)$ and $\lambda \sim G$ be the univariate model as in (3.14). The minimum Bayes risk is

$$R^*(G) = \inf_t \int E_\lambda(t(Y) - \lambda)^2 dG(\lambda) = \int E_\lambda(t^*_G(Y) - \lambda)^2 dG(\lambda),$$

where t^*_G is the oracle separable rule in (3.16). By Theorem 5 of [45], for $j \geq j_1 \equiv \max(\inf\{j\colon \log n_j \geq 2/p\}, j_0)$, there exists a universal constant M such that

$$\Big\{n_j^{-1} E_{\boldsymbol{\beta}_{[j]}} \Big\| \widehat{\boldsymbol{\beta}}_{[j]}/\sigma_N - \boldsymbol{\beta}_{[j]}/\sigma_N \Big\|^2 \Big\}^{1/2} - \Big\{R^*(G_{[j]})\Big\}^{1/2} \leq M\zeta_{n_j}(\log n_j)^{3/2}, \quad (3.40)$$

where $G_{[j]}(u) = n_j^{-1}\sum_{k=1}^{n_j} I(\beta_{jk}/\sigma_N \leq u)$ is the empirical distribution of the standardized vector $\boldsymbol{\beta}_{[j]}/\sigma_N$, and

$$\zeta_{n_j} = \max\Big\{\sqrt{2\log n_j}, \{n_j^{1/p}\sqrt{\log n_j}\mu_p^w(G_{[j]})\}^{p/(2+2p)}\Big\}\sqrt{\frac{\log n_j}{n_j}} \quad (3.41)$$

with $\mu_p^w(G) = \{\sup_{x>0} x^p \int_{|u|>x} G(du)\}^{1/p}$ as the p-th weak moment of a distribution G. By the definition of the weak moment and the Besov norm (3.8),

$$\mu_p^w(G_{[j]}) \leq \Big(\frac{1}{n_j}\sum_{k=1}^{n_j}\Big|\frac{\beta_{jk}}{\sigma_N}\Big|^p\Big)^{1/p} \leq Cn_j^{-(\alpha+1/2)}/\sigma_N. \quad (3.42)$$

From (3.41), it is easy to see

$$\zeta^2_{n_j} \leq \frac{2(\log n_j)^2}{n_j} + (\log n_j)^{1+p/(2+2p)}\Big(\frac{\mu_p^w(G_{[j]})}{n_j}\Big)^{p/(1+p)}. \quad (3.43)$$

By (3.40) and the inequality $2ab \leq a^2+b^2$, for any positive constant ξ_0,

$$\begin{aligned}\sum_{j=j_1}^J A_j &= \sigma_N^2 \sum_{j=j_1}^J n_j\Big\{n_j^{-1}E_{\boldsymbol{\beta}_{[j]}}\Big\|\widehat{\boldsymbol{\beta}}_{[j]}/\sigma_N - \boldsymbol{\beta}_{[j]}/\sigma_N\Big\|^2 - R^*(G_{[j]})\Big\} \\ &\leq \sigma_N^2\sum_{j=j_1}^J n_j\Big\{M\zeta_{n_j}(\log n_j)^{3/2}\Big\}\Big\{2\sqrt{R^*(G_{[j]})} + M\zeta_{n_j}(\log n_j)^{3/2}\Big\} \\ &\leq \sigma_N^2\sum_{j=j_1}^J n_j\Big\{M\Big(1+\frac{1}{\xi_0}\Big)(\log n_j)^3\zeta^2_{n_j} + \xi_0 R^*(G_{[j]})\Big\}. \quad (3.44)\end{aligned}$$

With (3.42), (3.44) and the upper bound (3.43), we have the following bound

$$\begin{aligned}\sum_{j=j_1}^J A_j &\leq M\sigma_N^2\Big(1+\frac{1}{\xi_0}\Big)\sum_{j=j_1}^J n_j\Big\{\frac{(\log n_j)^5}{n_j} + (\log n_j)^{4+\frac{p}{2+2p}}\Big(\frac{\mu_p^w(G_{[j]})}{n_j}\Big)^{\frac{p}{1+p}}\Big\} \\ &\quad +\sigma_N^2\xi_0\sum_{j=j_1}^J n_j R^*(G_{[j]}) \\ &\leq M\sigma_N^2\Big(1+\frac{1}{\xi_0}\Big)\sum_{j=j_1}^J n_j r_p\Big(n_j, \frac{Cn_j^{-(\alpha+1/2)}}{\sigma_N}\Big) \\ &\quad +\sigma_N^2\xi_0\sum_{j=j_1}^J n_j R^*(G_{[j]}), \quad (3.45)\end{aligned}$$

where $r_p(n,D) = (\log n)^5/n + (\log n)^{4+p/(2+2p)}(D/n)^{p/(1+p)}$.

We need to bound the rate of the second term in (3.45). This is done as follows. From (3.8), $\|\boldsymbol{\beta}_N\|^{\alpha}_{p,q} \le C$ if and only if for each $j \ge j_0$, $\|\boldsymbol{\beta}_{[j]}\|_p \equiv (n_j^{-1}\sum_{k=1}^{n_j}|\beta_{jk}|^p)^{1/p} \le C_j$ where C_j satisfy $(\sum C_j^q w_j^q)^{1/q} \le C$ with $w_j = 2^{-j(\alpha+1/2)}$,

$$\begin{aligned}
\mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)) &\equiv \inf_{\widehat{\boldsymbol{\beta}}_N}\sup_{\boldsymbol{\beta}_N\in B^{\alpha}_{p,q}(C)} R^{(N)}(\widehat{\boldsymbol{\beta}}_N,\boldsymbol{\beta}_N)\\
&= \inf_{\widehat{\boldsymbol{\beta}}_N}\sup_{(\sum C_j^q w_j^q)^{1/q}\le C}\sup_{\|\boldsymbol{\beta}_{[j]}\|_p\le C_j}\sum_{j=j_0}^{J} R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]},\boldsymbol{\beta}_{[j]})\\
&\ge \sup_{(\sum C_j^q w_j^q)^{1/q}\le C}\sum_{j=j_0}^{J}\inf_{\widehat{\boldsymbol{\beta}}_{[j]}}\sup_{\|\boldsymbol{\beta}_{[j]}\|_p\le C_j} R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]},\boldsymbol{\beta}_{[j]}). \qquad (3.46)
\end{aligned}$$

Moreover, by the minimax theory,

$$\inf_{\widehat{\boldsymbol{\beta}}_{[j]}}\sup_{\|\boldsymbol{\beta}_{[j]}\|_p\le C_j} R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]},\boldsymbol{\beta}_{[j]}) = \sigma_N^2\sup_{\|\boldsymbol{\beta}_{[j]}\|_p\le C_j}\Big\{n_j R^*(G_{[j]})(1+o(1))\Big\}. \qquad (3.47)$$

By (3.46) and (3.47), there exists some generic constant M so that

$$\begin{aligned}
\sigma_N^2\sup_{\boldsymbol{\beta}_N\in B^{\alpha}_{p,q}(C)}\sum_{j=j_0}^{J} n_j R^*(G_{[j]}) &= \sigma_N^2\sup_{(\sum C_j^q w_j^q)^{1/q}\le C}\sup_{\|\boldsymbol{\beta}_{[j]}\|_p\le C_j}\sum_{j=j_0}^{J} n_j R^*(G_{[j]})\\
&\le M\mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)). \qquad (3.48)
\end{aligned}$$

By the definition of minimax risk,

$$\begin{aligned}
\mathscr{R}(B^{\alpha}_{p,q}(C)) &= \inf_{\widehat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)} E\sum_{j=j_0}^{\infty}\sum_{k=1}^{n_j}(\widehat{\beta}_{jk}-\beta_{jk})^2\\
&\le \inf_{\widehat{\boldsymbol{\beta}}}\sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)} E\sum_{j=j_0}^{J}\sum_{k=1}^{n_j}(\widehat{\beta}_{jk}-\beta_{jk})^2 + \sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)}\sum_{j>J}\beta_{jk}^2\\
&\le \mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)) + \sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)}\sum_{j>J}\sum_{k}\beta_{jk}^2.
\end{aligned}$$

So that we have

$$\mathscr{R}(B^{\alpha}_{p,q}(C)) - \sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)}\sum_{j>J}\sum_{k}\beta_{jk}^2 \le \mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)) \le \mathscr{R}(B^{\alpha}_{p,q}(C)). \qquad (3.49)$$

By (3.10) of [76], $\mathscr{R}(B^{\alpha}_{p,q}(C)) \asymp N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}$. Moreover, by Hölder inequality, we have $\sup_{\boldsymbol{\beta}\in B^{\alpha}_{p,q}(C)}\sum_{j>J}\sum_k\beta_{jk}^2 \asymp N^{-2(\alpha+1/2-1/p)}$. When $\alpha^2/(\alpha +$

$1/2) > 1/p - 1/2$, the second term in the left hand side of (3.49) is negligible. Thus we have the minimax risk rate

$$\mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)) \asymp N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}. \tag{3.50}$$

Actually we have shown stronger result that $\mathscr{R}^{(N)}(B^{\alpha}_{p,q}(C)) = (1+o(1))\mathscr{R}(B^{\alpha}_{p,q}(C))$ uniformly for all Besov balls. Combining (3.45), (3.48) and (3.50) together, we have

$$\begin{aligned} & \sup_{\boldsymbol{\beta}_N \in B^{\alpha}_{p,q}(C)} \sum_{j=j_1}^{J} A_j \\ \leq \; & M\sigma_N^2\Big(1+\frac{1}{\xi_0}\Big)\sum_{j=j_1}^{J} n_j r_p\Big(n_j, \frac{Cn_j^{-(\alpha+1/2)}}{\sigma_N}\Big) \\ & + M\xi_0 N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}. \end{aligned} \tag{3.51}$$

Denote $(x_{jk}, \theta_{jk}) = (y_{jk}, \beta_{jk})/\sigma_N$ as the standardization with the unit variance. Then,

$$\begin{aligned} \sum_{j=j_0}^{j_1-1} A_j \; & \leq \; \sigma_N^2 \sum_{j=j_0}^{j_1-1} E_{\boldsymbol{\beta}_{[j]}} \Big\| \widehat{\boldsymbol{\beta}}_{[j]}/\sigma_N - \boldsymbol{\beta}_{[j]}/\sigma_N \Big\|^2 \\ & = \; \sigma_N^2 \sum_{j=j_0}^{j_1-1} E_{\boldsymbol{\theta}_{[j]}} \Big\| \boldsymbol{x}_{[j]} + \frac{f'_{\widehat{G}_{[j]}}(\boldsymbol{x}_{[j]})}{f_{\widehat{G}_{[j]}}(\boldsymbol{x}_{[j]}) \vee \rho_{n_j}} - \boldsymbol{\theta}_{[j]} \Big\|^2, \end{aligned} \tag{3.52}$$

where $f_{\widehat{G}_{[j]}}(x) = \int \varphi(x-u)\widehat{G}_{[j]}(du)$ and ρ_{n_j} is as in (3.22). Let $\widetilde{L}(\rho) = \sqrt{-\log(2\pi\rho^2)}$. By (3.22), and the fact that for any G, $|f'_G(x)|/(f_G(x) \vee \rho) \leq \widetilde{L}(\rho)$ when $0 < \rho < (2\pi e)^{-1/2}$ [45], there exists an constant M such that

$$\sum_{j=j_0}^{j_1-1} E_{\boldsymbol{\theta}_{[j]}} \Big\| \boldsymbol{x}_{[j]} + \frac{f'_{\widehat{G}_{[j]}}(\boldsymbol{x}_{[j]})}{f_{\widehat{G}_{[j]}}(\boldsymbol{x}_{[j]}) \vee \rho_{n_j}} - \boldsymbol{\theta}_{[j]} \Big\|^2 \leq \sum_{j=j_0}^{j_1-1} 2n_j(1+\widetilde{L}^2(\rho_{n_j})) \leq M \sum_{j=j_0}^{j_1-1} n_j \log n_j \tag{3.53}$$

In view of (3.52) and (3.53),

$$\sum_{j=j_0}^{j_1-1} A_j \leq M\sigma_N^2 \sum_{j=j_0}^{j_1-1} n_j \log n_j. \tag{3.54}$$

We arrive the oracle inequality (3.31) by combining (3.51) and (3.54) together.

□

Proof of Theorem 3.2. By (3.50) in the proof of Theorem 3.1, we only need to show that the left hand side of (3.33) is of smaller order than $N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}$.

Let δ be a constant such that $0 < \delta < p/[(\alpha+1/2)(1+p)]$ and $\gamma = (\alpha + 3/2)p/(1+p) - 1$. We set the critical index j_1 as

$$j_1 = \max(\lfloor \gamma^{-1}(p/(1+p) - 1/(\alpha+1/2) + \delta)\log_2(C/\sigma_N)\rfloor + 1, j_0), \qquad (3.55)$$

where $\lfloor x \rfloor$ stands for the largest integer no larger than x. As that in the proof of Theorem 3.1, we denote $A_j = R^{(N)}(\widehat{\boldsymbol{\beta}}_{[j]}, \boldsymbol{\beta}_{[j]}) - R^{(N,*)}(\boldsymbol{\beta}_{[j]})$ as the regret at the j-th level. We first show that the when cut off at the resolution level $j_1 - 1$, the sum of regret is of smaller than $N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}$. By (3.54) and (3.55), there exists a constant M such that

$$\begin{aligned}
\sum_{j=j_0}^{j_1-1} A_j &\le M\sigma_N^2 \sum_{j=j_0}^{j_1-1} n_j \log n_j \\
&\le M\sigma_N^2 (j_1+1)2^{j_1-1}\log 2^{j_1} \\
&\le M\sigma_N^2 j_1(j_1+1)\Big(\frac{C}{\sigma_N}\Big)^{\gamma^{-1}(p/(1+p)-1/(\alpha+1/2)+\delta)}.
\end{aligned}$$

Since $0 < \delta < p/[(\alpha+1/2)(1+p)]$, simple algebraic computation gives that $\gamma^{-1}(p/(1+p) - 1/(\alpha+1/2) + \delta < 1/(\alpha+1/2)$. Thus

$$\sum_{j=j_0}^{j_1-1} A_j \le o(1)N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}. \qquad (3.56)$$

From the (3.42) and the proof of Theorem 3.1, there exists a sequence of constants ξ_N such that

$$\begin{aligned}
\sum_{j=j_1}^{J} A_j &\le M\sigma_N^2 \sum_{j=j_1}^{J} n_j \Big(1 + \frac{1}{\xi_N}\Big)\Big\{\frac{(\log n_j)^5}{n_j} + (\log n_j)^{4+p/(2+2p)} \\
&\quad (Cn_j^{-(\alpha+3/2)}/\sigma_N)^{p/(1+p)}\Big\} + M\xi_N N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)} \\
&= M\sigma_N^2 \sum_{j=j_1}^{J}\Big(1 + \frac{1}{\xi_N}\Big)\Big\{(\log n_j)^5 + (\log n_j)^{4+p/(2+2p)} n_j^{-\gamma}\Big(\frac{C}{\sigma_N}\Big)^{p/(1+p)}\Big\} \\
&\quad + M\xi_N N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}. \qquad (3.57)
\end{aligned}$$

In view of the choice of γ and j_1 in (3.55), when $j \ge j_1$,

$$n_j^{-\gamma}\Big(\frac{C}{\sigma_N}\Big)^{p/(1+p)} \le \Big(\frac{C}{\sigma_N}\Big)^{1/(\alpha+1/2)-\delta}. \qquad (3.58)$$

Combining (3.57) and (3.58) leads to that

$$\begin{aligned}\sum_{j=j_1}^{J} A_j &\leq M\sigma_N^2 \sum_{j=j_1}^{J}\Big(1+\frac{1}{\xi_N}\Big)\Big\{(\log n_j)^5 + (\log n_j)^{4+p/(2+2p)}\Big(\frac{C}{\sigma_N}\Big)^{1/(\alpha+1/2)-\delta}\Big\}\\ &\quad + M\xi_N N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}\\ &\leq M\sigma_N^2 J\Big(1+\frac{1}{\xi_N}\Big)\Big\{(\log N)^5 + (\log N)^{4+p/(2+2p)}\Big(\frac{C}{\sigma_N}\Big)^{1/(\alpha+1/2)-\delta}\Big\}\\ &\quad + M\xi_N N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}. \end{aligned} \tag{3.59}$$

We pick ξ_N satisfying $\xi_N \to 0$ and $\xi_N N^{\delta/2} \to \infty$. Then, under the calibration $\sigma_N^2 = \sigma_N^2$,

$$M\xi_N \to 0, \quad \Big(1+\frac{1}{\xi_N}\Big)\Big(\frac{C}{\sigma_N}\Big)^{-\delta} \to 0.$$

Thus, from (3.59), the order of sum of regret at the resolution levels from j_1 to J is also smaller than $N^{-\alpha/(\alpha+1/2)}C^{1/(\alpha+1/2)}$. We arrive the uniform ideal adaptation (3.33). □

Proof of Theorem 3.3. By (3.33) in Theorem 3.2, under condition (3.10),

$$\sup_{\boldsymbol{\beta}_N \in B_{p,q}^{\alpha}(C)} R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N) \leq o(1)\mathscr{R}^{(N)}(B_{p,q}^{\alpha}(C)) + \sup_{\boldsymbol{\beta}_N \in B_{p,q}^{\alpha}(C)} R^{(N,*)}(\boldsymbol{\beta}_N). \tag{3.60}$$

where $R^{(N)}(\widehat{\boldsymbol{\beta}}_N, \boldsymbol{\beta}_N)$ and $R^{(N,*)}(\boldsymbol{\beta}_N)$ are as in (3.28) and (3.30) respectively.

By (3.11) of [76], for all Besov balls $B_{p,q}^{\alpha}(C)$,

$$\sup_{\boldsymbol{\beta}_N \in B_{p,q}^{\alpha}(C)} R^{(N,*)}(\boldsymbol{\beta}_N) \leq (1+o(1))\mathscr{R}(B_{p,q}^{\alpha}(C)). \tag{3.61}$$

In the proof of Theorem 3.1, we have shown that

$$\mathscr{R}^{(N)}(B_{p,q}^{\alpha}(C)) = (1+o(1))\mathscr{R}(B_{p,q}^{\alpha}(C))$$

under (3.10). This fact with (3.61) demonstrate that

$$\sup_{\boldsymbol{\beta}_N \in B_{p,q}^{\alpha}(C)} R^{(N,*)}(\boldsymbol{\beta}_N) \leq (1+o(1))\mathscr{R}^{(N)}(B_{p,q}^{\alpha}(C)). \tag{3.62}$$

The exactly adaptive minimaxity (3.34) follows from (3.60) and (3.62). □

Chapter 4
A Penalized Linear Unbiased Selection Algorithm for Generalized Linear Model

4.1 Introduction

By discovering important relevant variables, variable selection can improve upon the prediction accuracy and interpretability of a statistical model. Classical variable selection procedures such as AIC, BIC, and C_p essentially impose a penalty on loss based on the number of selected variables. Using such model selection criterion, we need to evaluate each candidate model and pick up the best one. This is computationally infeasible with even moderately high-dimensional data. Hence, regularization techniques are used to fulfill continuous selection. The LASSO [68] method minimizes the square loss function with the ℓ_1 penalty on the parameters in a linear regression model. Due to the singularity of the ℓ_1 penalty at the origin, the LASSO has variable selection feature of shrinking some coefficients exactly to zero [25]. Under the same paradigm, the penalized negative log-likelihood approach with the ℓ_1 penalty is used to select variables in generalized linear models [50]. Fan and Li [34] advocated that a good penalty should result in an estimator with unbiasedness, sparsity and continuity. They formulated the smoothly clipped absolute deviation (SCAD) penalty which provides certain oracle properties. The SCAD enjoys the oracle property in term of selection accuracy and estimation efficiency when the regularization parameter is appropriately chosen. However, the computation of the SCAD is challenging because of its concavity over $(0, \infty)$. Recently, motivated by alleviating the degree of concavity of the SCAD, Zhang

[78] proposed the minimax concave (MC) penalty. The MC minimizes the maximum concavity among all penalties satisfying an unbiasedness condition. Both the SCAD and MC are spline quadratic functions which are singular at the origin and concave over $(0, \infty)$.

Implementation of concave penalization methodologies demands an efficient algorithm to compute the selector at different penalty levels, or even better, a path of solutions encompassing a suitable range of penalty levels. This is crucial since the "best" penalty level is typically data driven. The computation of the LASSO paths is relatively friendly since the ℓ_1 penalty is convex. In the linear regression case, efficient algorithms have been developed for the exact computation of the LASSO path [55, 56, 29]. The computation of the concave penalty path is much more difficult since the penalized loss might be non-convex. Inspired by algorithms for the LASSO, Zhang [78] proposed the penalized linear unbiased selection (PLUS) algorithm to compute the solution paths of possibly non-convex penalized least squares. The PLUS algorithm assumes that the penalty function is a quadratic spline in $[0, \infty)$ so that the LASSO, SCAD and MC methods are included. The PLUS continuously tracks a path in certain main branch of solution graph of possibly multiple local minimizers. It computes multiple local minimizers at an individual penalty level by continuously tracing a path of critical values of the penalized loss at different penalty levels. This special computational strategy of the PLUS enables it to efficiently generate a solution path of concave penalized least squares.

In contrast to the great advance achieved in linear model, computation of penalized selection and estimation in the generalized linear model is considerably less developed. In this area, several algorithms for approximating a (local) minimizer at an individual pre-selected penalty level have been developed. This type of algorithms includes the local quadratic approximation (LQA) [34], the minorize-maximize (MM) algorithm [43] and the local linear approximation (LLA) [86] for the SCAD method, and the CLG algorithm [37] for large scale ℓ_1-penalized logistic model. Park and Hastie [57] and Zhao and Yu [84] proposed path approximation algorithms for the minimization of the ℓ_1-penalized negative log-likelihood. However, as far as we are aware, a path approximation algorithm for the concave-penalized negative log-likelihood does not exist.

In this chapter we propose the generalized PLUS (GPLUS) algorithm to compute the paths of concave-penalized generalized linear model. The GPLUS retains the same mechanism of the PLUS to find the multiple local minimizers and the same assumption that the penalty is a concave quadratic spline function. Being different with linear model, the paths of generalized linear model are not piecewise linear. Our strategy is approximating the nonlinear paths with end-to-end short linear segments. The length of each segment controls the overall accuracy of the path. The new algorithm works in a stagewise fashion: in each iteration, the paths traverse along the current direction with a small step. We prove that under suitable regularity conditions, the computed paths converge to the true paths over a certain range of penalty levels. In the simulation study, we put emphasis on applying the GPLUS algorithm to the penalized logistic regression model because of its importance in data classification and prediction. Interestingly, depending on whether the minimization problem is convex and whether the solution paths are piecewise linear, the computational strategies of path following algorithms are different. We shall discuss the relationship among several existing algorithms in this chapter.

The remaining part of this chapter is organized as follows. In Section 4.3, we discuss concave-penalized negative log-likelihood approach for variable selection. In Section 4.3 we present the GPLUS algorithm to compute the concave penalization method and discuss its relationship with other existing path following algorithms. We show numerical examples with both simulated and real data in Section 4.4. Section 4.5 is a discussion and quick summary. The mathematical proof is contained in Section 4.6.

4.2 The Concave Penalization Method

We first introduce some notations which are used throughout this chapter. Consider a dataset $\boldsymbol{Z} \equiv \{(\boldsymbol{x}_i, y_i)_{i=1}^n\}$ containing n identically and independently distributed observations, where y_i are response variables and $\boldsymbol{x}_i \in \mathbb{R}^p$ are predictors. The n by p design matrix is $\boldsymbol{X} \equiv (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$ where $\boldsymbol{x}^j$ is the j-th variable.

In generalized linear model, y_i depends on $\boldsymbol{x}_i$ through a linear combination $\boldsymbol{x}_i^T \boldsymbol{\beta}$ where $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)^T$. The regression coefficient $\beta_j = 0$ means the j-th variable do not influence the response. Model selection aims to locate those variables $\boldsymbol{x}^j$ with nonzero β_j. Given $\boldsymbol{x}_i$ and y_i, the log-likelihood is $\ell_i(\boldsymbol{\beta}, \phi) \equiv \ell_i(\boldsymbol{x}_i^T \boldsymbol{\beta}, y_i, \phi)$ where ϕ is a dispersion parameter. In logistic regression, no dispersion parameter ϕ exists. In linear regression, the estimation of ϕ has no influence on the estimation of $\boldsymbol{\beta}$. Therefore, the penalized negative log-likelihood approach does not penalize ϕ, and the log-likelihood can be written as $\ell_i(\boldsymbol{\beta}) \equiv \ell_i(\boldsymbol{\beta}, \phi)$. The regularized estimates are given by

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} \left\{ \psi(\boldsymbol{\beta}) + \sum_{j=1}^{p} \rho(|\beta_j|; \lambda) \right\}, \tag{4.1}$$

where $\psi(\boldsymbol{\beta}) \equiv -\frac{1}{n} \sum_{i=1}^n \ell_i(\boldsymbol{\beta})$ is the negative log-likelihood and $\rho(t; \lambda)$ is a penalty function indexed by regularization parameter $\lambda \geq 0$.

The LASSO method uses the ℓ_1 penalty $\rho(t; \lambda) = \lambda t$ with $t \geq 0$. The ℓ_1 penalty is the only member generating continuous and sparse estimation among ℓ_α ($\alpha > 0$) family of penalties, but it will result in estimation bias. Some recent research on the LASSO consistency show that, due to the bias, strong conditions are required for selection consistency under the ℓ_1 penalty in the linear regression model [51, 83, 81].

In the earlier studies on the effect of the bias of more general penalized estimators on estimation efficiency, Fan and Li [34] suggested using a penalty function which keeps a constant beyond certain level so that the bias of sufficiently large coefficient is nearly removed. They carefully formulated the SCAD penalty

$$\rho(t;\lambda) = \lambda \int_0^t \min\left\{1, \frac{(\gamma - x/\lambda)_+}{(\gamma-1)}\right\} \mathrm{d}x, \quad \gamma > 2, \tag{4.2}$$

as a variable selector to realize their advocation and showed that the SCAD performs as well as the oracle procedure in terms of selecting the correct subset model and estimating the true nonzero coefficients.

The SCAD penalty (4.2) satisfies the constraints

$$\dot{\rho}(t;\lambda) = 0 \text{ for } t \geq \gamma\lambda, \quad \dot{\rho}(0+;\lambda) = \lambda, \tag{4.3}$$

where $\dot{\rho}(t;\lambda) \equiv (\partial/\partial t)\rho(t;\lambda)$. However, these constraints will result in concave penalties, or equivalently non-convex penalized negative log-likelihood. Zhang [78] pointed out that the degree of concavity of the penalized negative log-likelihood considerably influences the computational complexity of path. There, Zhang further proposed the MC

$$\rho(t;\lambda) = \lambda \int_0^t \left(1 - \frac{x}{\gamma\lambda}\right)_+ \mathrm{d}x, \tag{4.4}$$

as the minimizer of the maximum concavity

$$\kappa(\rho;\lambda) \equiv \sup_{t>0} \Big\{ -\ddot{\rho}(t;\lambda) \Big\}, \quad \ddot{\rho}(t;\lambda) \equiv (\partial/\partial^2 t)\rho(t;\lambda), \tag{4.5}$$

among all penalty functions satisfying the constraints (4.3). The penalty function has selection features if $\dot{\rho}(0+;\lambda) > 0$. The second part of (4.3) standardizes the index λ so that it has the interpretation as the threshold for β_j for standardized designs with $\|\boldsymbol{x}^j\|^2/n = 1$. Fan and Li [34] pointed out that the first part of (4.3) allows nearly unbiased estimation for β_j with large absolute values. Thus, (4.3) is called the unbiased selection conditions [79]. Being the minimizer of the maximum concavity among all penalty functions satisfying the unbiased selection conditions, the MC method retains the convexity of the penalized negative log-likelihood in (4.1) to the greatest extent under constraints (4.3). Conversely, given the maximum concavity $\kappa(\rho;\lambda)$, the MC provides the smallest $\gamma\lambda$ which is the left end of unbiased selection region $(\gamma\lambda, \infty)$. The maximum concavity

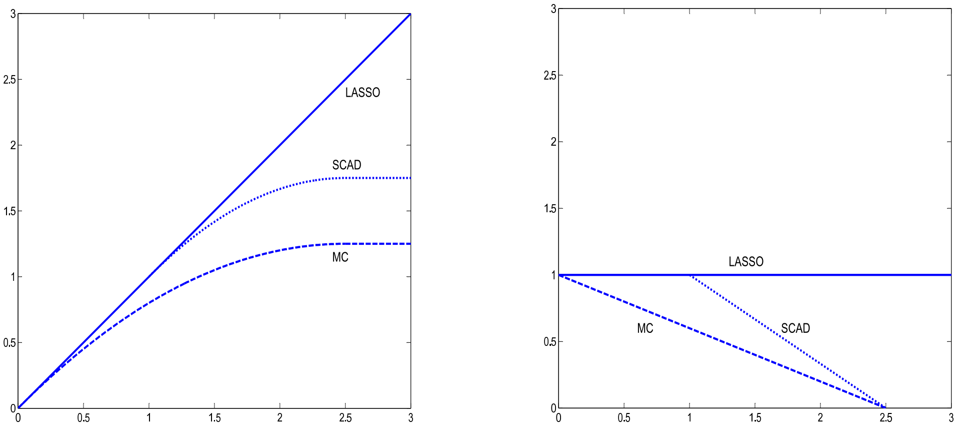

Figure 4.1: The penalty functions for the LASSO (solid), SCAD (dotted) and MC (dashed) with $\gamma = 2.5$. Left: the penalties $\rho(t)$; right: their derivatives $\dot{\rho}(t)$.

$\kappa(\rho;\lambda) = 1/(\gamma - 1)$ and $1/\gamma$ for the SCAD and MC, respectively. Hence, the tuning parameter γ in the SCAD and MC regulates the computational complexity of the solution paths via controlling the maximum concavity $\kappa(\rho;\lambda)$.

Theoretical investigation shows that the concave-penalized least squares methods possess selection consistency and oracle efficiency properties under much weaker conditions than the ℓ_1 penalized methods do. Such desirable properties are expected to extend to the generalized linear models including the logistic regression. However, due to the singularity and concavity of the penalty function, minimization of concave-penalized negative log-likelihood is still commonly viewed as a computationally challenging problem. Computational difficulties of (4.1) such as multiple local minimizers will arise.

Mathematically, the ℓ_1, the SCAD and the minimax concave penalties described above are all special cases of more general penalties of the form $\rho(t;\lambda) = \lambda^2\rho(t/\lambda)$, where $\rho(t)$ is an increasing quadratic spline in $[0,\infty)$. Such $\rho(t)$ have piecewise linear, continuous and nonnegative derivative $\dot{\rho}(t)$ for $t \geq 0$.

$$\rho(t;\lambda) = \lambda^2\rho(t/\lambda), \quad \dot{\rho}(t) \equiv \frac{d}{dt}\rho(t) = \sum_{i=1}^{m}(u_i - v_i t)I(t_i \leq t \leq t_{i+1}) \qquad (4.6)$$

with knots $0 = t_1 < t_2 < \cdots < t_m = \gamma$ satisfying $u_i - v_i t_{i+1} = u_{i+1} - v_{i+1}t_{i+1}, i = 1, \ldots, m$. We set $u_1 = 1$, $u_m = v_m = 0$ and $t_{m+1} = \infty$ so that condition (4.3) holds. The penalty class (4.6) includes the ℓ_1 penalty as a special member with

$m = 1$ and $\dot{\rho}(t) = 1$, the MC with $m = 2$ and $\dot{\rho}(t) = (1 - t/\gamma)_+$, and the SCAD penalty with $m = 3$ and $\dot{\rho}(t) = \min(1, (\gamma - t)_+/(\gamma - 1))$. Figure 4.1 shows these three penalties with their derivatives for $\gamma = 2.5$.

4.3 The GPLUS Algorithm

In this section, we describe the GPLUS algorithm in detail. The GPLUS is designed to approximate the solution paths of optimization problem (4.1) where the penalty function $\rho(t; \lambda)$ is a quadratic spline as in (4.6). Throughout the section we refrain from specifying the concrete form of $\psi(\boldsymbol{\beta})$ since the derivations apply equally to any $\psi(\boldsymbol{\beta})$ with continuous first two derivatives with respect to $\boldsymbol{\beta}$. We divide this section into 3 subsections to cover the Karush-Kuhn-Tucker type condition, the GPLUS algorithm, and the comparison of several existing path following algorithm.

4.3.1 The Karush-Kuhn-Tucker type condition and the PLUS algorithm

With penalty of the form (4.6), the Karush-Kuhn-Tucker type condition of optimization problem (4.1) is

$$\begin{cases} \dot{\psi}_j(\boldsymbol{\beta}(\lambda)) + \lambda\,\mathrm{sgn}(\beta_j(\lambda))\,\dot{\rho}(|\beta_j(\lambda)|/\lambda) = 0, & \text{if } \beta_j(\lambda) \neq 0, \\ |\dot{\psi}_j(\boldsymbol{\beta}(\lambda))| \leq \lambda, & \text{if } \beta_j(\lambda) = 0. \end{cases} \tag{4.7}$$

where $\dot{\boldsymbol{\psi}} \in \mathbb{R}^p$ is the gradient vector of ψ. In order to solve the minimization problem (4.1), essentially we need to trace the solutions of (4.7) as λ varies.

Under the scale transformation $\tau \equiv 1/\lambda$ and $\boldsymbol{b}(\tau) \equiv \boldsymbol{\beta}(\lambda)/\lambda$, (4.7) becomes to be

$$\begin{cases} \tau\dot{\psi}_j(\boldsymbol{b}(\tau)/\tau) + \mathrm{sgn}(b_j(\tau))\,\dot{\rho}(|b_j(\tau)|) = 0, & \text{if } b_j(\tau) \neq 0, \\ \tau|\dot{\psi}_j(\boldsymbol{b}(\tau)/\tau)| \leq 1, & \text{if } b_j(\tau) = 0. \end{cases} \tag{4.8}$$

Condition (4.8) is an equivalent version of (4.7), while constant 1 in the inequality constraint will provide convenience in the derivation of algorithm. Therefore, in the remainder of this article, we work with (4.8) instead of (4.7).

Define

$$u(i) \equiv u_{|i|}, \quad v(i) \equiv v_{|i|}, \quad t(i) \equiv \begin{cases} t_i, & \text{if } 0 < i \leq m+1, \\ -t_{|i|+1}, & \text{if } -m \leq i \leq 0, \end{cases} \tag{4.9}$$

where the u_i, v_i and t_i are as in (4.6). Let $\boldsymbol{\eta} \in \{-m, \ldots, m\}^p$ be a p-indicator such that

$$t(\eta_j) \leq b_j(\tau) \leq t(\eta_j + 1), \quad j = 1, \ldots, p. \tag{4.10}$$

In other words, $\boldsymbol{\eta}$ represents the penalty intervals of $\boldsymbol{b}(\tau)$.

When $t(\eta_j) \leq b_j(\tau) \leq t(\eta_j + 1)$, by (4.6), we have $\operatorname{sgn}(b_j(\tau))\dot{\rho}(|b_j(\tau)|) = \operatorname{sgn}(\eta_j)u(\eta_j) - b_j(\tau)v(\eta_j)$. We rewrite (4.8) in more explicit form: define

$$\begin{aligned} S(\boldsymbol{\eta}) \quad \equiv \quad & \text{all } \boldsymbol{y} \oplus \boldsymbol{b} \text{ satisfying} \\ & \begin{cases} \tau\dot{\psi}_j(\boldsymbol{b}(\tau)/\tau) + \operatorname{sgn}(\eta_j)u(\eta_j) - b_j(\tau)v(\eta_j) = 0, & \text{if } \eta_j \neq 0, \\ -1 \leq \tau\dot{\psi}_j(\boldsymbol{b}(\tau)/\tau) \leq 1, & \text{if } \eta_j = 0, \\ t(\eta_j) \leq b_j(\tau) \leq t(\eta_j + 1), & \text{if } \eta_j \neq 0, \\ b_j(\tau) = 0, & \text{if } \eta_j = 0. \end{cases} \end{aligned} \tag{4.11}$$

(4.8) holds iff (4.11) holds for certain $\boldsymbol{\eta}$. For each $\boldsymbol{\eta}$, since (4.11) has p equations and p pairs of parallel inequalities, $S(\boldsymbol{\eta})$ are p-dimensional blocks living in $\mathbb{R}^{2p}$. Due to the continuity of $\dot{\rho}(t) = (d/dt)\rho(t)$ in t by (4.6) and that of $\dot{\psi}_j$ in both $\boldsymbol{y}$ and $\boldsymbol{b}$, the solutions of (4.11) are identical in the intersection of any pair of $S(\boldsymbol{\eta})$ with adjacent $\boldsymbol{\eta}$. Moreover, the p-dimensional interiors of different $S(\boldsymbol{\eta})$ are disjoint in view of the constraints on $\boldsymbol{b}$ of (4.11). Thus, the union of all the p-dimensional blocks $S(\boldsymbol{\eta})$ forms a continuous p-dimensional surface $S \equiv \cup\big\{S(\boldsymbol{\eta})\colon\ \boldsymbol{\eta} \in \{-m, \ldots, m\}^p\big\}$ in $\mathbb{R}^{2p}$. Given data $\boldsymbol{y}$, the solution set of (4.8) is the intersection of this p-surface S and the p-subspace $\{\boldsymbol{y} \oplus \boldsymbol{b}\colon\ \boldsymbol{b} \in \mathbb{R}^p\}$.

Let $\boldsymbol{P}_{\boldsymbol{\eta}}$ be the projection matrix $\boldsymbol{P}_{\boldsymbol{\eta}}\boldsymbol{b} = (b_j\colon\ \eta_j \neq 0)^T$, the first equation of (4.11) could be written in matrix notation

$$\tau \boldsymbol{P}_{\boldsymbol{\eta}}\big(\dot{\boldsymbol{s}}(\boldsymbol{b}(\tau)/\tau)\big) + \boldsymbol{P}_{\boldsymbol{\eta}}\big(\mathrm{sgn}(\boldsymbol{\eta})u(\boldsymbol{\eta}) - \boldsymbol{b}(\tau)v(\boldsymbol{\eta})\big) = 0, \tag{4.12}$$

where the multiplication in the second parentheses is componentwise. Let $\ddot{\boldsymbol{\Psi}} \in \mathbb{R}^{p\times p}$ be the Hessian matrix of ψ and denote

$$\begin{aligned}
\boldsymbol{Q}(\boldsymbol{b}(\tau),\tau) &= \boldsymbol{P}_{\boldsymbol{\eta}}\ddot{\boldsymbol{\Psi}}\big(\boldsymbol{b}(\tau)/\tau\big)\boldsymbol{P}_{\boldsymbol{\eta}}^T - \mathrm{diag}\big(v(\eta_j),\, \eta_j \neq 0\big), && (4.13)\\
\boldsymbol{w}(\boldsymbol{b}(\tau),\tau) &= \boldsymbol{P}_{\boldsymbol{\eta}}\big(\ddot{\boldsymbol{\Psi}}(\boldsymbol{b}(\tau)/\tau)\boldsymbol{b}(\tau)/\tau - \dot{\boldsymbol{s}}(\boldsymbol{b}(\tau)/\tau)\big). && (4.14)
\end{aligned}$$

Taking differentiation of (4.12) with respect to τ, we get the differentiation form of the KKT equation

$$\boldsymbol{Q}(\boldsymbol{b}(\tau),\tau)\,\boldsymbol{P}_{\boldsymbol{\eta}}\,\boldsymbol{s}(\boldsymbol{b}(\tau),\tau) = \boldsymbol{w}(\boldsymbol{b}(\tau),\tau), \quad \eta_j = 0 \Rightarrow s_j = 0. \tag{4.15}$$

where $\boldsymbol{s}(\boldsymbol{b}(\tau),\tau) = (d/d\tau)\boldsymbol{b}(\tau)$ is the local "slope" of $\boldsymbol{b}(\tau)$.

To get insights into the GPLUS algorithm described in next subsection, we give a quick review of the PLUS algorithm for the linear regression model where $\psi(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2/(2n)$ is taken to be the squared loss. It is easy to show that $\boldsymbol{Q}(\boldsymbol{b}(\tau),\tau) = \boldsymbol{P}_{\boldsymbol{\eta}}(\boldsymbol{X}^T\boldsymbol{X}/n)\boldsymbol{P}_{\boldsymbol{\eta}}^T - \mathrm{diag}(v(\eta_j),\, \eta_j \neq 0)$ and $\boldsymbol{w}(\boldsymbol{b}(\tau),\tau) = \boldsymbol{P}_{\boldsymbol{\eta}}(\boldsymbol{X}^T\boldsymbol{y}/n)$. This implies that the slope $\boldsymbol{s}(\boldsymbol{b}(\tau),\tau)$ is constant in each block $\boldsymbol{\eta}$, because when $\boldsymbol{\eta}$ does not change, $\boldsymbol{P}_{\boldsymbol{\eta}}, \boldsymbol{Q}(\boldsymbol{b}(\tau),\tau)$ and $\boldsymbol{w}(\boldsymbol{b}(\tau),\tau)$ will not change either. Hence it indicates that the solution paths $\boldsymbol{b}(\tau)$ are piecewise linear in τ. The piecewise linearity will greatly facilitate the computation of entire trajectories: as long as we find all the turning points, all values in between are obtained by linear interpolation. In linear regression model, almost everywhere in $\boldsymbol{X}$ and γ, the solution set of (4.8) is composed of a main branch and separate loops. The main branch is piecewise linear, begins with $\boldsymbol{b} = 0$, and ends with least squares solution satisfying $\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}(\tau^{(k^*)})/\tau^{(k^*)}) = 0$. The PLUS algorithm traces the main branch of the solution paths by computing one line segment between two turning points in each step.

4.3.2 The GPLUS algorithm

In generalized linear models, the paths are not piecewise linear. Our strategy is discretely approximating the nonlinear paths by many end-to-end short line segments.

The GPLUS procedure works roughly as follows. We start with $\boldsymbol{b}^{(0)} = 0$, and find the largest possible value $\tau = \tau^{(0)}$ at which all b_j are zero. In the k-th iteration, with one endpoint $\boldsymbol{b}^{(k-1)}$, we compute a second endpoint of the k-th piece of segment. We firstly find the index $\boldsymbol{\eta}^{(k)}$ which indicates the block where the k-th piece of segment lives. Starting from $\boldsymbol{b}^{(k-1)}$, the paths proceed in a direction $\boldsymbol{s}^{(k)}$ with the step size $\Delta^{(k)}$ until $\boldsymbol{b}^{(k)}$. $\boldsymbol{s}^{(k)}$ is decided by all the equation constraints in (4.8). The step size $\Delta^{(k)}$ is designed up to be a pre-determined constant Δ and to make sure that each piece of segment is wholly contained in one block. In other words, we "cut" the path exactly at the block boundary when it is going to enter into a new block. The main updating rules are

$$\tau^{(k)} = \tau^{(k-1)} + \xi^{(k)}\Delta^{(k)}, \quad \boldsymbol{b}^{(k)} = \boldsymbol{b}^{(k-1)} + (\tau^{(k)} - \tau^{(k-1)})\boldsymbol{s}^{(k)},$$

where $\xi^{(k)} = \pm 1$ characterizes whether the paths go back or forth with respect to τ in each iteration. In the GPLUS algorithm, the value of τ may not be monotone increasing. When $\xi^{(k)} = -1$, τ will decrease so that the multiple local minimizers are obtained (same τ, different $\boldsymbol{b}$). Once we obtain all the turning points $(\boldsymbol{b}^{(k)}, \tau^{(k)})$, the paths are given by linear interpolation

$$\widehat{\boldsymbol{\beta}}(\lambda) = \widehat{\boldsymbol{b}}(\tau)/\tau, \quad \widehat{\boldsymbol{b}}(\tau) = \frac{\tau^{(k)} - \tau}{\tau^{(k)} - \tau^{(k-1)}}\boldsymbol{b}^{(k-1)} + \frac{\tau - \tau^{(k-1)}}{\tau^{(k)} - \tau^{(k-1)}}\boldsymbol{b}^{(k)}. \tag{4.16}$$

We summarize the GPLUS algorithm in the following syllabus and explain in details next.

Initialization: Compute $\boldsymbol{\eta}^{(0)}$, $\tau^{(0)}$ and $\boldsymbol{b}^{(0)}$. Set $k = 1$.

Iterations:

1. Compute the block index $\boldsymbol{\eta}^{(k)}$.
2. Compute the proceeding direction vector $\boldsymbol{s}^{(k)}$.
3. Compute the direction indicator $\xi^{(k)} = 1$ or -1.
4. Compute the step length $\Delta^{(k)}$.
5. Compute $\tau^{(k)}$ and $\boldsymbol{b}^{(k)}$. Increase k by one, $k \leftarrow k + 1$.

Output: $\boldsymbol{\eta}^{(k)}$, $\boldsymbol{b}^{(k)}$, $\tau^{(k)}, k = 0, 1, \ldots, k^*$.

- **Initialization**

We initialize the GPLUS algorithm with

$$\boldsymbol{\eta}^{(0)} = 0, \quad \tau^{(0)} = 1/\max_j |\dot{\psi}_j^{(0)}|, \quad \boldsymbol{b}^{(0)} = 0, \tag{4.17}$$

where $\dot{\psi}_j^{(0)} = \dot{\psi}_j(\boldsymbol{b})|_{\boldsymbol{b}=0}$. In view of the inequalities in the KKT condition (4.8), the initial segment is $\boldsymbol{b}(\tau) = 0$ for all $0 \leq \tau \leq \tau^{(0)}$. We note that $\lambda^{(0)} = 1/\tau^{(0)}$ is the smallest value of λ that makes all b_j zero, $j = 1, \ldots, p$.

- **Iterations**

In the k-th iteration, we compute $\boldsymbol{\eta}^{(k)}$, $\boldsymbol{s}^{(k)}$, $\xi^{(k)}$, $\Delta^{(k)}$, $\tau^{(k)}$ and $\boldsymbol{b}^{(k)}$ in sequence based on $\boldsymbol{\eta}^{(k-1)}$, $\tau^{(k-1)}$ and $\boldsymbol{b}^{(k-1)}$. As mentioned in the syllabus above, each iteration is divided into 5 steps.

Step 1: compute $\boldsymbol{\eta}^{(k)}$. Denote $\dot{\boldsymbol{\psi}}^{(k-1)} \equiv \dot{\boldsymbol{\psi}}(\boldsymbol{b}^{(k-1)}/\tau^{(k-1)})$ and $\ddot{\boldsymbol{\Psi}}^{(k-1)} \equiv \ddot{\boldsymbol{\Psi}}(\boldsymbol{b}^{(k-1)}/\tau^{(k-1)})$ as the gradient and Hessian of $\psi(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{b}^{(k-1)}/\tau^{(k-1)}$. Let

$$\begin{aligned} C^{(k-1)} \equiv C_1^{(k-1)} \cup C_2^{(k-1)} \quad \equiv \quad & \Big\{ j\colon \, |b_j^{(k-1)}| \in \{t_1, \ldots, t_m\} \text{ with } \eta_j^{(k-1)} \neq 0 \Big\} \\ & \cup \Big\{ j\colon \, |\tau^{(k-1)} \dot{\psi}_j^{(k-1)}| \geq 1 \text{ with } \eta_j^{(k-1)} = 0 \Big\}. \end{aligned}$$

be the set of critical indices j of which b_j hits the boundary of the inequalities in (4.11) at $\tau = \tau^{(k-1)}$. Based on $\boldsymbol{\eta}^{(k-1)}$ and $C^{(k-1)}$, we compute next index vector

$\boldsymbol{\eta}^{(k)}$ according to the following rule,

$$\eta_j^{(k)} = \begin{cases} \eta_j^{(k-1)}, & \text{if } j \notin C^{(k-1)}, \\ \eta_j^{(k-1)} + 1 & \text{if } j \in C_1^{(k-1)} \text{ and } \xi^{(k-1)} s_j^{(k-1)} > 0, \\ \eta_j^{(k-1)} - 1, & \text{if } j \in C_1^{(k-1)} \text{ and } \xi^{(k-1)} s_j^{(k-1)} < 0, \\ 1, & \text{if } j \in C_2^{(k-1)} \text{ and } \tau^{(k-1)} \dot{\psi}_j^{(k-1)} \leq -1, \\ -1, & \text{if } j \in C_2^{(k-1)} \text{ and } \tau^{(k-1)} \dot{\psi}_j^{(k-1)} \geq 1. \end{cases} \tag{4.18}$$

Step 2: compute $\boldsymbol{s}^{(k)}$. Let $\boldsymbol{Q}^{(k)}$ and $\boldsymbol{w}^{(k)}$ be as defined in (4.13) and (4.14) but depend only on $\boldsymbol{b}^{(k-1)}$, $\tau^{(k-1)}$ and $\boldsymbol{\eta}^{(k)}$. More explicitly,

$$\begin{aligned} \boldsymbol{Q}^{(k)} &= \boldsymbol{P}_{\boldsymbol{\eta}^{(k)}} \ddot{\boldsymbol{\Psi}}^{(k-1)} \boldsymbol{P}_{\boldsymbol{\eta}^{(k)}}^T - \operatorname{diag}\big(v(\eta_j^{(k)}),\, \eta_j^{(k)} \neq 0\big), & (4.19) \\ \boldsymbol{w}^{(k)} &= \boldsymbol{P}_{\boldsymbol{\eta}^{(k)}} \big(\ddot{\boldsymbol{\Psi}}^{(k-1)} \boldsymbol{b}^{(k-1)} / \tau^{(k-1)} - \dot{\boldsymbol{\psi}}^{(k-1)}\big), & (4.20) \end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{\eta}^{(k)}}$ is the projection matrix such that $\boldsymbol{P}_{\boldsymbol{\eta}^{(k)}} \boldsymbol{z} = (z_j \colon \eta_j^{(k)} \neq 0)^T$. The direction vector $\boldsymbol{s}^{(k)}$ is determined by the equation

$$\boldsymbol{Q}^{(k)} \boldsymbol{P}_{\boldsymbol{\eta}^{(k)}} \boldsymbol{s}^{(k)} = \boldsymbol{w}^{(k)}, \quad \eta_j^{(k)} = 0 \Rightarrow s_j^{(k)} = 0, \tag{4.21}$$

where $\boldsymbol{Q}^{(k)}$ and $\boldsymbol{w}^{(k)}$ are defined in (4.19) and (4.20). The progress direction $\boldsymbol{s}^{(k)}$ can be view as compromise among the currently active covariates.

Step 3: compute $\xi^{(k)}$. Given $\boldsymbol{\eta}^{(k)}$ and $\boldsymbol{s}^{(k)}$, we pick the direction indicator $\xi^{(k)} = 1$ or -1 which make $\boldsymbol{s}^{(k)}$ indeed carry the k-th segment of the paths from $S(\boldsymbol{\eta}^{(k-1)})$ to $S(\boldsymbol{\eta}^{(k)})$. Formally, the definition of $\xi^{(k)}$ is $\xi^{(k)} \equiv \operatorname{sgn}(\tau^{(k)} - \tau^{(k-1)})$. It decides whether the paths go ahead ($\xi^{(k)} = 1$) or back ($\xi^{(k)} = -1$) in current iteration. Since $\tau^{(k)}$ is unknown at this moment, we utilize $\boldsymbol{\eta}^{(k)}$ and $\boldsymbol{s}^{(k)}$ to characterize it.

If $C^{(k-1)}$ is empty, that is, $\boldsymbol{\eta}^{(k-1)} = \boldsymbol{\eta}^{(k)}$, then the $(k-1)$-th and k-th segments are in the same block. $\xi^{(k)}$ is given by

$$\xi^{(k)} = \begin{cases} -\xi^{(k-1)}, & \text{if } \operatorname{sgn}(\boldsymbol{s}^{(k)}) \operatorname{sgn}(\boldsymbol{s}^{(k-1)}) \in \{-1, 0\}^p, \\ \xi^{(k-1)}, & \text{otherwise.} \end{cases} \tag{4.22}$$

In another word, when $\boldsymbol{\eta}^{(k-1)} = \boldsymbol{\eta}^{(k)}$, if at least one pair of $\boldsymbol{s}^{(k)}$ and $\boldsymbol{s}^{(k-1)}$ have same sign, the PLUS paths will keep the same direction as previous step. Otherwise the paths will turn around.

If $C^{(k-1)}$ is non-empty, it amounts to verify the following set of conditions

$$\begin{cases} \xi^{(k)}(\eta_j^{(k)} - \eta_j^{(k-1)})s_j^{(k)} \geq 0, & \text{if } \eta_j^{(k-1)} \neq \eta_j^{(k)} \neq 0, \\ \xi^{(k)}\eta_j^{(k-1)}\frac{d}{d\tau}(\tau^{(k-1)}\dot{\psi}_j^{(k-1)}) \geq 0, & \text{if } \eta_j^{(k-1)} \neq \eta_j^{(k)} = 0. \end{cases} \tag{4.23}$$

Step 4: compute $\Delta^{(k)}$. Let $\Delta^{(k)} \equiv |\tau^{(k)} - \tau^{(k-1)}|$ be the length of the k-th segment of the paths measured in τ. If we fix the step length to be some constant Δ in each iteration, it will happen that two endpoints of certain segment are located in different blocks. In such cases, b_j may directly change from positive value to negative without staying at zero, or vice verse. To avoid such jumps, we cut the path exactly on the boundary when the crossing is going to happen. The allowed maximum step size $\Delta_j^{(k)}$ of the j-th coordinate is

$$\Delta_j^{(k)} = \begin{cases} \xi^{(k)}\{t(\eta_j^{(k)} + 1) - b_j^{(k-1)}\}/s_j^{(k)}, & \text{if } \eta_j^{(k)} \neq 0 \text{ and } \xi^{(k)}s_j^{(k)} > 0, \\ \xi^{(k)}\{t(\eta_j^{(k)}) - b_j^{(k-1)}\}/s_j^{(k)}, & \text{if } \eta_j^{(k)} \neq 0 \text{ and } \xi^{(k)}s_j^{(k)} < 0, \\ \infty, & \text{if } \eta_j^{(k)} = 0. \end{cases} \tag{4.24}$$

Finally, $\Delta^{(k)}$ is given by

$$\Delta^{(k)} = \min\{\Delta,\, \Delta_j^{(k)},\, 1 \leq j \leq p\} \tag{4.25}$$

Step 5: compute $\tau^{(k)}$ and $\boldsymbol{b}^{(k)}$.

$$\tau^{(k)} = \tau^{(k-1)} + \xi^{(k)}\Delta^{(k)}, \quad \boldsymbol{b}^{(k)} = \boldsymbol{b}^{(k-1)} + (\tau^{(k)} - \tau^{(k-1)})\boldsymbol{s}^{(k)}, \tag{4.26}$$

Remark 4.1. *Since the end-to-end line segments computed are not the exact solutions to (4.8), it may happen that when $b_j^{(k-1)}$ re-hit the knot $t = 0$, $|\tau^{(k-1)}\dot{\psi}_j^{(k-1)}| > 1$ and $|\tau^{(k)}\dot{\psi}_j^{(k)}| > 1$. Consequently, according to the rule (4.18), we have $\eta_j^{(k-1)} \neq 0$, $\eta_j^{(k)} = 0$ and $\eta_j^{(k+1)} \neq 0$. Thus, the j-th variable is excluded from and included into the model alternatively during the consecutive iterations. To prevent the selection from such oscillation, once we observe that*

$$\eta_j^{(k-1)} \neq 0, \quad b_j^{(k-1)} = 0, \quad |\tau^{(k-1)}\dot{\psi}_j^{(k-1)}| > 1, \tag{4.27}$$

we shall exclude the j-th variable from the model until the value of $|\tau\dot{\psi}_j|$ is strictly smaller than 1 and then goes beyond 1 again.

Remark 4.2. *The GPLUS algorithm works based on the notion that in each block, tiny amount of departure from the true paths at very beginning stage will not cause dramatic error when the approximation proceeds. This stability property results from the continuity conditions on ψ and ρ. The parameter Δ controls how close the algorithm approximates the path. A smaller step size leads to a closer approximation.*

Remark 4.3. *In the GPLUS algorithm, we trace the gradients to select new variables and use Hessian matrix to find the proceeding directions of selected variables simultaneously. As is in (4.21), we only need to compute the inverse sub-Hessian matrix of selected variables. Hence the computation of matrix inverse is efficient in sparse model.*

Theorem 4.1. *Let $\boldsymbol{b}(\tau)$ be the solution of (4.8) and $\widehat{\boldsymbol{b}}(\tau)$ be the computed paths as in (4.16). We use the notation $\boldsymbol{b} \in \boldsymbol{\eta}$ to denote the condition (4.10), that is, $\boldsymbol{b}$ lives in the block indexed by $\boldsymbol{\eta}$. Assume that $\boldsymbol{b}(\tau)$ is second differentiable with respect to τ. For the block indexed by $\boldsymbol{\eta}$, if the following conditions hold:*

(i) Let k_0 be the smallest integer such that $\{\widehat{\boldsymbol{b}}(\tau^{(k_0)}), \boldsymbol{b}(\tau^{(k_0)})\} \in \boldsymbol{\eta}$, the initial error $\|\widehat{\boldsymbol{b}}(\tau^{(k_0)}) - \boldsymbol{b}(\tau^{(k_0)})\| \to 0$ as $\Delta \to 0$.

(ii) For the block $\boldsymbol{\eta}$, there exists constants δ and M_1 such that when $\|\boldsymbol{b} - \boldsymbol{b}(\tau)\| \leq \delta$, the Lipschitz condition $\|\boldsymbol{s}(\boldsymbol{b}, \tau) - \boldsymbol{s}(\boldsymbol{b}(\tau), \tau)\| \leq M_1 \|\boldsymbol{b} - \boldsymbol{b}(\tau)\|$ holds where $\boldsymbol{s}(\cdot, \tau)$ is as in (4.15) and $\{\boldsymbol{b}, \boldsymbol{b}(\tau)\} \in \boldsymbol{\eta}$.

(iii) There exists a constant M_2 such that $\sup_{\{\tau :\, \boldsymbol{b}(\tau) \in \boldsymbol{\eta}\}} \|\frac{d}{d\tau}\boldsymbol{s}(\boldsymbol{b}(\tau), \tau)\| \leq M_2$.

(iv) There exists a constant M_3 such that $\sup_{\Delta > 0} \sum_{\{j :\, \widehat{\boldsymbol{b}}(\tau^{(j)}) \in \boldsymbol{\eta}\}} \Delta^{(j+1)} < M_3$.

Then, when $\{\widehat{\boldsymbol{b}}(\tau^{(k)}), \boldsymbol{b}(\tau^{(k)})\} \in \boldsymbol{\eta}$, $\|\widehat{\boldsymbol{b}}(\tau^{(k)}) - \boldsymbol{b}(\tau^{(k)})\| \to 0$ as $\Delta \to 0$.

4.3.3 Discussion of path following algorithms

From a unified point of view, all the paths algorithms aim at the same task: tracing the solution paths of a set of the KKT equations (4.8). However, depending on whether the optimization problem is convex and whether the solution paths are piecewise linear, the computational strategies are pretty different.

As mentioned earlier, the computation of the LASSO paths is relatively friendly since the ℓ_1-penalized least squares is convex and the LASSO paths are piecewise linear. Efficient algorithms that give the entire Lasso paths have been established, namely, the homotopy method [55, 56] and similarly the LARS algorithm [29]. The LARS provides a nice geometrical interpretation of these methods: in the ℓ_1-penalized least squares problem, the selected variables share the same absolute correlation with current fitting residuals $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Thus, in each step, the LARS proceeds equiangularly between the selected variables (variables with nonzero estimated coefficient), that is, along the "least angle direction", until next variable enters or one of selected variables is removed.

The PLUS algorithm essentially generalizes the LARS to compute the concave-penalized least squares. Since there are multiple phases of concave penalty, the paths are much more complex but still piecewise linear. The GPLUS algorithm applies the computational strategy of the PLUS to the generalized linear model in a pretty direct manner. However, because of the nonlinearity of the generalized linear model path, the procedures are more deliberate.

Motivated by the LARS, Park and Hastie [57] introduced an efficient path following algorithm for ℓ_1-penalized generalized linear model. Their algorithm computes the entire nonlinear solution paths by using the predictor-corrector method of convex optimization. In each iteration, with certain carefully chosen λ at which the set of non-zero coefficients changes, the corrector step finds accurate minimizer (4.1) of the convex objective function based on a good starting estimator provided by the predictor step. Their method uses end-to-end linear segments of moderate size to approximate the nonlinear paths and yields exact order of the variable selection. However, Park and Hastie's predictor-corrector method is not suitable for concave penalty since the corrector step is difficult to implement in the non-convex minimization problem. That explains why we proceed with tiny steps in the GPLUS algorithm.

We summarize the relationship between the aforementioned algorithms in the following chart.

$$\begin{tikzcd}
\text{LARS} \arrow[r, "\text{concave penalty}"] \arrow[d, swap, "\text{GLM}"] & \text{PLUS} \arrow[d, "\text{GLM}"] \\
\text{Park and Hastie}'\text{s} \arrow[r, swap, "\text{concave penalty}"] & \text{GPLUS}
\end{tikzcd}$$

All four algorithms can be viewed as moderately greedy forward stepwise procedures whose progress direction is determined by compromise among the currently selected variables. Hence in each iteration, we update all coordinates simultaneously. Zhao and Yu [84] proposed the BLASSO algorithm which update only one coordinate each time. Therefore their method avoids the matrix inversion. The BLASSO is designed to approximate the paths of any ℓ_1-penalized convex loss function by accommodating the backward steps into the forward stagewise fitting. The backward step in the BLASSO can remove the selected irrelevant variables.

4.4 Numerical Experiments

4.4.1 Linear regression

In this experiment, we compare the selection accuracy of the LASSO, SCAD and MC methods in linear model

$$\boldsymbol{y} = \sum_{j=1}^{p} \beta_j \boldsymbol{x}^j + \boldsymbol{\varepsilon}, \tag{4.28}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the response vector, $\boldsymbol{x}^j = (x_{1j}, \ldots, x_{nj})^T \in \mathbb{R}^n, j = 1, \ldots, p$, are p predictors, $\boldsymbol{\beta} \equiv (\beta_1, \ldots, \beta_p)^T$ are regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ are noises. Hereafter throughout this paper, we denote $A^o \equiv \{j : \beta_j \neq 0\}$ as the set of variables contribute to the model, $\widehat{A} \equiv \{j : \widehat{\beta}_j \neq 0\}$ as the set of selected variables, and $d^o \equiv |A^o| = \#\{j : \beta_j \neq 0\}$ is the size of A^o.

Our design is randomly generated as described below to guarantee a fair amount of correlation among the covariates. We generate an $n \times p$ random design $\boldsymbol{X}$ with each observation $\boldsymbol{x}_i$ following an AR(1) model. In detail, we generate the observations one by one independently. For the i-th observation, the first covariate x_{i1} is sampled from the standard normal distribution. For $j = 2, \ldots, p$, the j-th feature is generated according to the AR(1) model $x_{ij} = \rho x_{i,j-1} + e_{ij}$ where e_{ij} are independent $N(0, \varepsilon^2)$ random variables. We set $\rho^2 + \varepsilon^2 = 1$ so that each variable has unit variance, i.e., $\mathrm{Var}(x_{ij}) = 1, j = 1, \ldots, p$. Finally we normalize $\boldsymbol{X}$ so that each column has mean zero and ℓ_2 norm $\sqrt{n}$. In this experiment, our dimension setting is $(n, p) = (300, 200)$.

We evaluate selection performance for low correlation $\rho = 0.25$ and high correlation $\rho = 0.75$ in Table 4.1 and 4.2, respectively. In each table, there are three measurements: $CS \equiv I\{\widehat{A} = A^o\}$ is the indicator of correct selection, $TM \equiv |\widehat{A} \backslash A^o| + |A^o \backslash \widehat{A}|$ is the total miss as the sum of false discovery and missing discovery, and k is the number of steps. All the results reported are based on 100 replications. In each replication, A^o, $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are randomly sampled. The response vector $\boldsymbol{y}$ is generated from model (4.28) where $\beta_j = \pm\beta^*$ for $j \in A^o$, $\beta_j = 0$ for $j \notin A^o$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}_n)$. The parameters here are $(\beta^*, \gamma) = (0.7, 3.7)$ where γ is the regularization parameter as in (4.2) and (4.4). The value $\gamma = 3.7$ is suggested by Fan and Li (2001). We present the results at four penalty levels: $\lambda = \hat{\sigma}\sqrt{a(\log p)/n}, a = 1, \ldots, 4$, where $\hat{\sigma}^2$ is the mean squared error with $n - p$ degrees of freedom in full rank design. Bold face entries indicate $P\{\widehat{A} = A^o\} \approx \overline{CS} > 0.5$.

As can be seen from Table 4.1, it is clear that the variable selection accuracy of the SCAD and MC dominate the LASSO. The superiority is overwhelming when $d^o = 20$ and 40. Especially, when $d^o = 40$, the LASSO only correctly identify the true variables at most once among 100 replications with four different penalty levels, while the MC still shows strong selection accuracy with $\lambda = \widehat{\sigma}\sqrt{2(\log p)/n}$ and $\widehat{\sigma}\sqrt{4(\log p)/n}$. In fact, the simulation and theoretical results in [78] show that the universal penalty level $\lambda = \widehat{\sigma}\sqrt{2(\log p)/n}$ is nearly the optimal choice

Table 4.1: Performance of the LASSO, SCAD and MC methods in linear regression based on 100 replications: $n = 300$, $p = 200$, $\beta^* = 0.7$, $\gamma = 3.7$. Each observation is generated from an AR(1) model $x_{ij} = \rho x_{i,j-1} + e_{ij}$ with low correlation $\rho = 0.25$, $e_{ij} \sim N(0, \varepsilon^2)$, and $\rho^2 + \varepsilon^2 = 1$

		$d^o = 10$			$d^o = 20$			$d^o = 40$		
$\lambda/\widehat{\sigma}$		LASSO	SCAD	MC	LASSO	SCAD	MC	LASSO	SCAD	MC
$\sqrt{(\log p)/n}$ =0.1329	$\overline{CS}$	0.00	0.03	0.03	0.00	0.09	0.09	0.00	0.13	0.13
	$\overline{TM}$	5.89	3.66	3.49	8.45	2.92	2.85	13.08	2.19	2.14
	$\bar{k}$	17	35	24	30	65	44	55	132	86
$\sqrt{2(\log p)/n}$ =0.1879	$\overline{CS}$	0.28	**0.75**	**0.75**	0.07	**0.83**	**0.84**	0.01	**0.85**	**0.90**
	$\overline{TM}$	1.26	0.30	0.30	2.63	0.21	0.19	7.77	0.16	0.11
	$\bar{k}$	12	27	17	24	53	32	49	112	66
$\sqrt{4(\log p)/n}$ =0.2658	$\overline{CS}$	**0.80**	**0.97**	**1.00**	0.46	**0.84**	**0.97**	0.01	0.13	**0.83**
	$\overline{TM}$	0.20	0.03	0.00	0.83	0.19	0.03	5.32	2.44	0.18
	$\bar{k}$	11	21	11	22	40	21	46	77	44
$\sqrt{8(\log p)/n}$ =0.3759	$\overline{CS}$	**0.93**	**0.93**	**1.00**	**0.53**	**0.51**	**0.80**	0.00	0.00	0.06
	$\overline{TM}$	0.07	0.07	0.00	0.65	0.66	0.25	6.81	6.85	4.46
	$\bar{k}$	11	14	11	21	27	21	41	51	39

Table 4.2: The comparison of selections where each observation of the design matrix $\boldsymbol{X}$ is generated from an AR(1) model with high correlation $\rho = 0.75$. Other settings are the same as those in Table 4.1

		$d^o = 10$			$d^o = 20$			$d^o = 40$		
$\lambda/\widehat{\sigma}$		LASSO	SCAD	MC	LASSO	SCAD	MC	LASSO	SCAD	MC
$\sqrt{(\log p)/n}$ =0.1329	$\overline{CS}$	0.00	0.20	0.20	0.00	0.25	0.25	0.00	0.09	0.17
	$\overline{TM}$	9.36	1.90	1.76	14.71	1.49	1.46	24.11	4.00	3.52
	$\bar{k}$	21	42	28	36	124	84	66	1544	884
$\sqrt{2(\log p)/n}$ =0.1879	$\overline{CS}$	0.01	**0.67**	**0.73**	0.00	0.20	0.31	0.00	0.00	0.00
	$\overline{TM}$	5.13	0.54	0.45	10.99	2.06	1.84	23.36	12.06	11.17
	$\bar{k}$	16	32	20	30	82	55	58	895	466
$\sqrt{4(\log p)/n}$ =0.2658	$\overline{CS}$	0.05	0.43	**0.60**	0.00	0.00	0.07	0.00	0.00	0.00
	$\overline{TM}$	3.23	1.09	0.84	9.06	4.98	3.64	23.67	20.44	18.69
	$\bar{k}$	13	23	13	26	57	32	49	494	254
$\sqrt{8(\log p)/n}$ =0.3759	$\overline{CS}$	0.11	0.13	0.32	0.00	0.00	0.00	0.00	0.00	0.00
	$\overline{TM}$	2.45	2.21	1.34	8.60	8.10	6.14	25.20	25.58	24.00
	$\bar{k}$	12	15	11	22	37	24	39	203	112

Table 4.3: Performance of the LASSO, SCAD and MC methods in linear regression based on 100 replications: $n = 300$, $p = 200$, $\beta^* = 1/2$, $\gamma = 3.7$. The design matrix $\boldsymbol{X}$ is generated by greedy sequential group sampling from a larger pool random matrix

$\lambda/\widehat{\sigma}$		$d^o = 10$			$d^o = 20$			$d^o = 40$		
		LASSO	SCAD	MCP	LASSO	SCAD	MCP	LASSO	SCAD	MCP
$\sqrt{(\log p)/n}$ =0.1329	$\overline{CS}$	0.00	0.05	0.04	0.00	0.05	0.07	0.00	0.06	0.13
	$\overline{TM}$	5.60	3.77	3.63	8.03	3.39	3.22	11.17	2.65	2.28
	$\bar{k}$	17	30	20	29	56	35	53	114	67
$\sqrt{2(\log p)/n}$ =0.1879	$\overline{CS}$	0.42	0.68	0.79	0.09	0.38	0.72	0.00	0.09	0.52
	$\overline{TM}$	1.03	0.37	0.25	2.64	0.98	0.33	6.73	3.76	0.78
	$\bar{k}$	12	21	11	24	40	22	47	78	44
$\sqrt{4(\log p)/n}$ =0.2658	$\overline{CS}$	0.87	0.87	0.95	0.31	0.35	0.64	0.00	0.01	0.09
	$\overline{TM}$	0.13	0.13	0.05	1.23	1.10	0.48	6.49	6.66	4.27
	$\bar{k}$	11	14	11	21	27	21	42	52	40
$\sqrt{8(\log p)/n}$ =0.3759	$\overline{CS}$	0.40	0.40	0.45	0.03	0.03	0.05	0.00	0.00	0.00
	$\overline{TM}$	1.11	1.11	1.06	4.72	4.73	4.83	13.32	13.43	14.01
	$\bar{k}$	10	10	10	17	17	16	31	32	29

for variable selection in linear model with $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ and the normalization $\|\boldsymbol{x}^j\|^2/n = 1$. Moreover, the computational complexity of the MC is competitive with the LASSO as demonstrated by the average number of steps $\bar{k}$.

In Table 4.2, we report the selection performances when each observation of the design matrix $\boldsymbol{X}$ is generated from an AR(1) model with high correlation $\rho = 0.75$. Other settings are the same as those in Table 4.1. As expected, with higher correlations among the variables, the computation of the SCAD and MC is more costly. Dramatic rise in the number of computation steps is observed when $d^o = 40$. Again, the average $\overline{CS}$ and $\overline{TM}$ over 100 replications demonstrates the superior performance of the concave methods in our simulation experiments. Especially, when $d^o = 20$ and 40, the LASSO fails to identify the true variables correctly in each replication, while the other two demonstrate considerable accuracy with proper penalty amount.

We present another set of simulation results in Table 4.3. In this simulation, we first generate a $n \times p^*$ pool random matrix $\boldsymbol{X}^{pl}$ with each cell iid standard normal random variable. We normalize $\boldsymbol{X}^{pl}$ so that each column has mean zero and

ℓ_2 norm $\sqrt{n}$. The design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ is generated by greedy sequential group sampling from $\boldsymbol{X}^{pl}$. In this experiment, $(n, p^*, p) = (300, 1000, 200)$. Each sampling group consists of 20 most correlated vectors from the remaining columns of $\boldsymbol{X}^{pl}$. For the m-th group, we sample from the remaining $1020 - 20m$ columns one member $\boldsymbol{x}_{20m-19}$ and 19 more to maximize the absolute correlation $|\boldsymbol{x}_j' \boldsymbol{x}_{20m-19}|/n$, $j = 20m - 18, \ldots, 20m$, $m = 1, \ldots, 10$. The design $\boldsymbol{X}$ is fixed throughout this experiment, the maximum absolute correlation between the columns is 0.2299. The selection results in Table 4.3 are similar to Table 4.1, it is clear that the variable selection accuracy of the MC is better than the other two methods. The superiority is overwhelming when $d^o = 20$ and 40. When $d^o = 40$, the LASSO fails to identify the correct set in every replication, while the MC still shows strong selection accuracy with $\lambda = \widehat{\sigma}\sqrt{2(\log p)/n}$.

4.4.2 Logistic regression

In this example, we assess the performance of the GPLUS algorithm for logistic regression model. In logistic regression model, we have a set of n independent pairs $\boldsymbol{x}_i$, y_i, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a p-vector of predictors for the i-th observation. Given $\boldsymbol{x}_i$, $y_i \in \{0, 1\}$ is the i-th binary response with probability of success

$$p_i \equiv P(y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}. \tag{4.29}$$

The loss function $\psi(\boldsymbol{\beta})$ is taken to be the negative log-likelihood

$$\psi(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n} \left(y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}))\right). \tag{4.30}$$

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p)$ be design matrix of size $n \times p$. The gradient and Hessian matrix are $\dot{\boldsymbol{\psi}}(\boldsymbol{\beta}) = -\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{p})/n$ and $\ddot{\boldsymbol{\Psi}}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}/n$ where $\boldsymbol{p} = (p_1, \ldots, p_n)^T$ and $\boldsymbol{W}$ is an $n \times n$ diagonal matrix of weights with i-th element $p_i(1 - p_i)$.

Each observation $\boldsymbol{x}_i$ is drawn independently from a multivariate normal distribution with zero mean and correlation $\rho^{|j-k|}$ between j-th and k-th entries with $\rho = 0.5$. The number of covariate predictors is $p = 100$. We generate a training design $\boldsymbol{X}$ with $n = 200$ observations. Again, $\boldsymbol{X}$ is normalized so that each column has mean zero and ℓ_2 norm $\sqrt{n}$. Throughout this experiment, $\boldsymbol{X}$ is fixed . All the results reported are based on 100 replications. In each replication, as the procedures in Experiment 1, A^o and $\boldsymbol{\beta}$ are sampled with $\beta_j = \pm\beta^*$ for $j \in A^o$, $\beta_j = 0$ for $j \notin A^o$. Then the response y is generated according to (4.29). Besides reporting the correct selection $\overline{CS}$ and total miss $\overline{TM}$, we compute the Kullback-Leibler divergence between the true distribution P and its estimation $\widehat{P}$,

$$\begin{aligned} KL(P, \widehat{P}) &= P(y=1)\log\frac{P(y=1)}{\widehat{P}(y=1)} + P(y=0)\log\frac{P(y=0)}{\widehat{P}(y=0)} \\ &= \boldsymbol{x}^T(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta})} + \log\frac{1+\exp(\boldsymbol{x}^T\widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta})}. \end{aligned} \qquad (4.31)$$

We use a Monte Carlo simulation to compute the Kullback-Leibler divergence (4.31).

We provide two sets of diagrams in Figure 4.2-4.5 with $d^o = 5$ and $d^o = 10$ respectively. Other parameters are $\beta^* = 1.25$, $\gamma = 16$ or 32 and step size $\Delta = 0.005$. We plot the $\overline{CS}$, $\overline{TM}$ and $\overline{KL}$ as functions of λ based on 100 replications. As can be seen from Figure 4.2-4.5, the SCAD and MC make considerable improvement over the LASSO in the sense of better selection accuracy, smaller total miss and Kullback-Leibler divergence. In fact, among the interval of λ plotted, the SCAD and MC overwhelmingly dominate the LASSO. Moreover, the performances of the MC are always a little bit better than those of the SCAD.

In Figure 4.6, we plot the solution for one replication with parameters $n = 200$, $p = 100$, $d^o = 5$, $\beta^* = 1.25$, $\gamma = 16$ and $\Delta = 0.005$. The solutions plotted are build up in 4000 steps. Middle (the SCAD) and right (the MC) panels are nearly indistinguishable from each other. All the three methods select five true variables in the early stage. In the first 4000 steps, the LASSO selects more noisy variables than the other two methods. An interesting phenomenon exhibited in Figure 4.6 is that, for the SCAD and MC paths, there exists an interval of λ during which

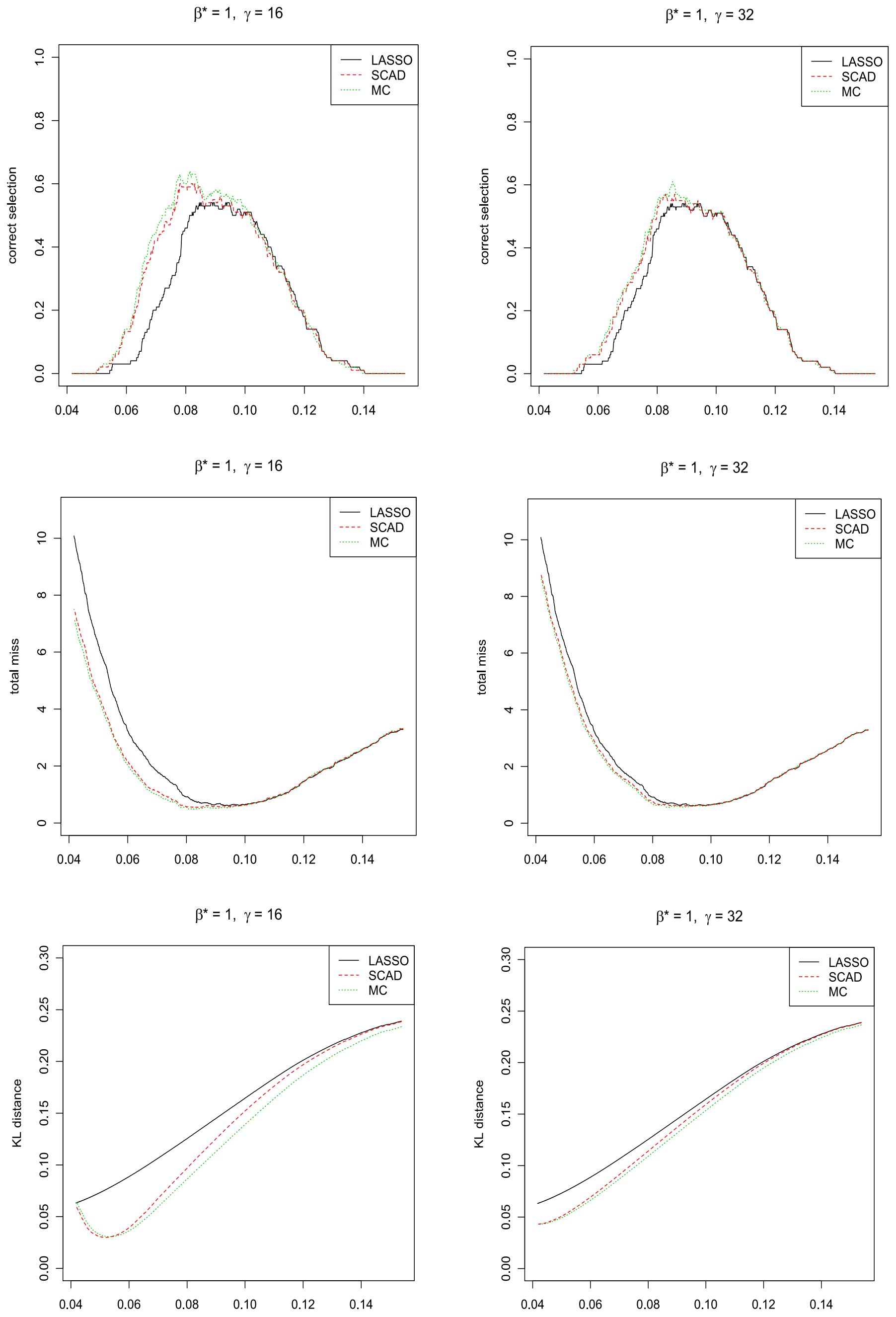

Figure 4.2: The comparison between the LASSO (black, solid), SCAD (red, dashed) and MC (green, dotted) in logistic regression based on 100 replications. The correct selection probability $\overline{CS}$, total miss $\overline{TM}$ and Kullback-Leibler divergence $\overline{KL}$ are plotted against the penalty level λ. Top panels: $\overline{CS}$; middle panels: $\overline{TM}$; bottom panels: $\overline{KL}$; left panels: $\gamma = 16$; right panels: $\gamma = 32$. Parameters: $n = 200$, $p = 100$, $d^o = 5$, $\beta^* = 1$ and $\Delta = 0.005$.

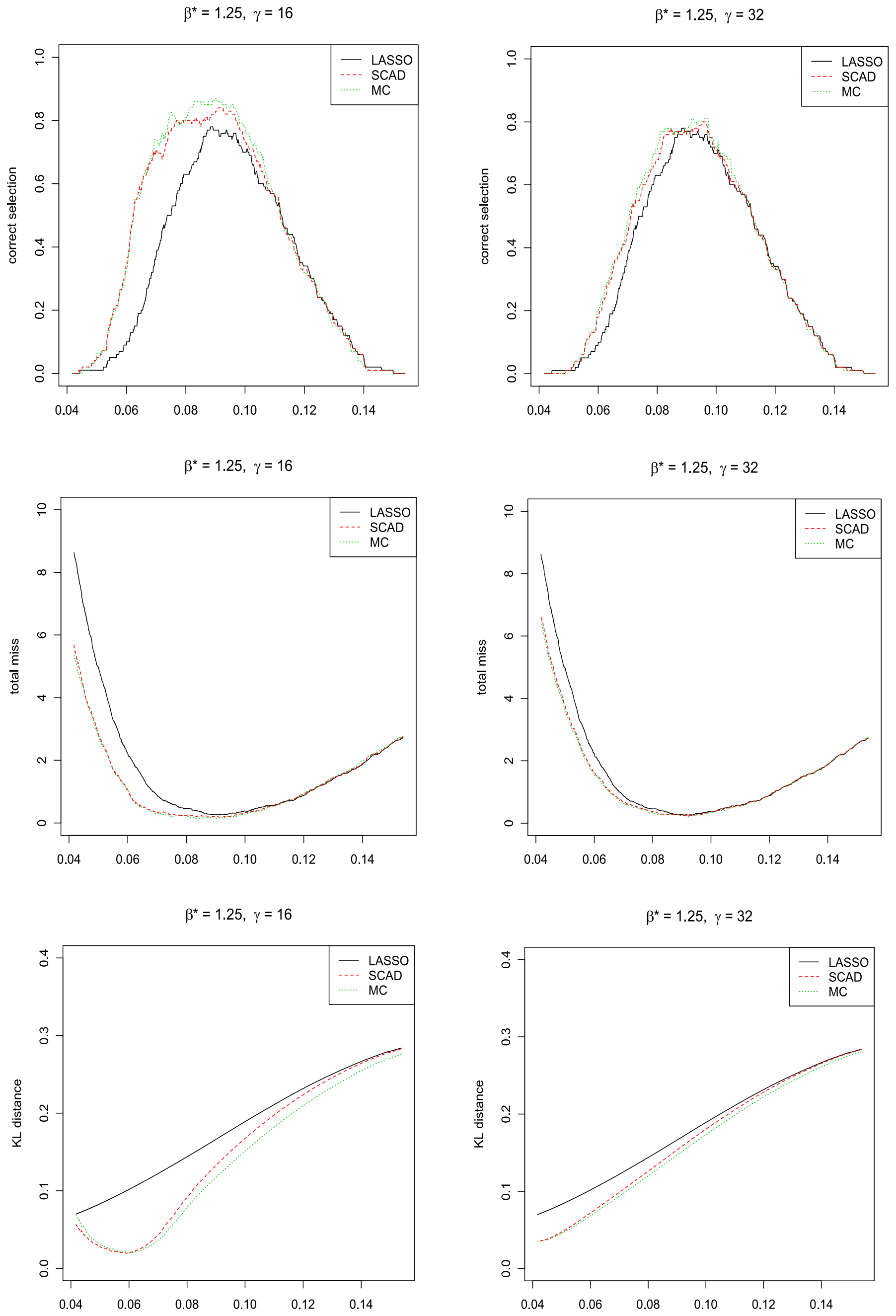

Figure 4.3: The comparison between the LASSO (black, solid), SCAD (red, dashed) and MC (green, dotted) in logistic regression based on 100 replications. The correct selection probability $\overline{CS}$, total miss $\overline{TM}$ and Kullback-Leibler divergence $\overline{KL}$ are plotted against the penalty level λ. Top panels: $\overline{CS}$; middle panels: $\overline{TM}$; bottom panels: $\overline{KL}$; left panels: $\gamma = 16$; right panels: $\gamma = 32$. Parameters: $n = 200$, $p = 100$, $d^o = 5$, $\beta^* = 1.25$ and $\Delta = 0.005$.

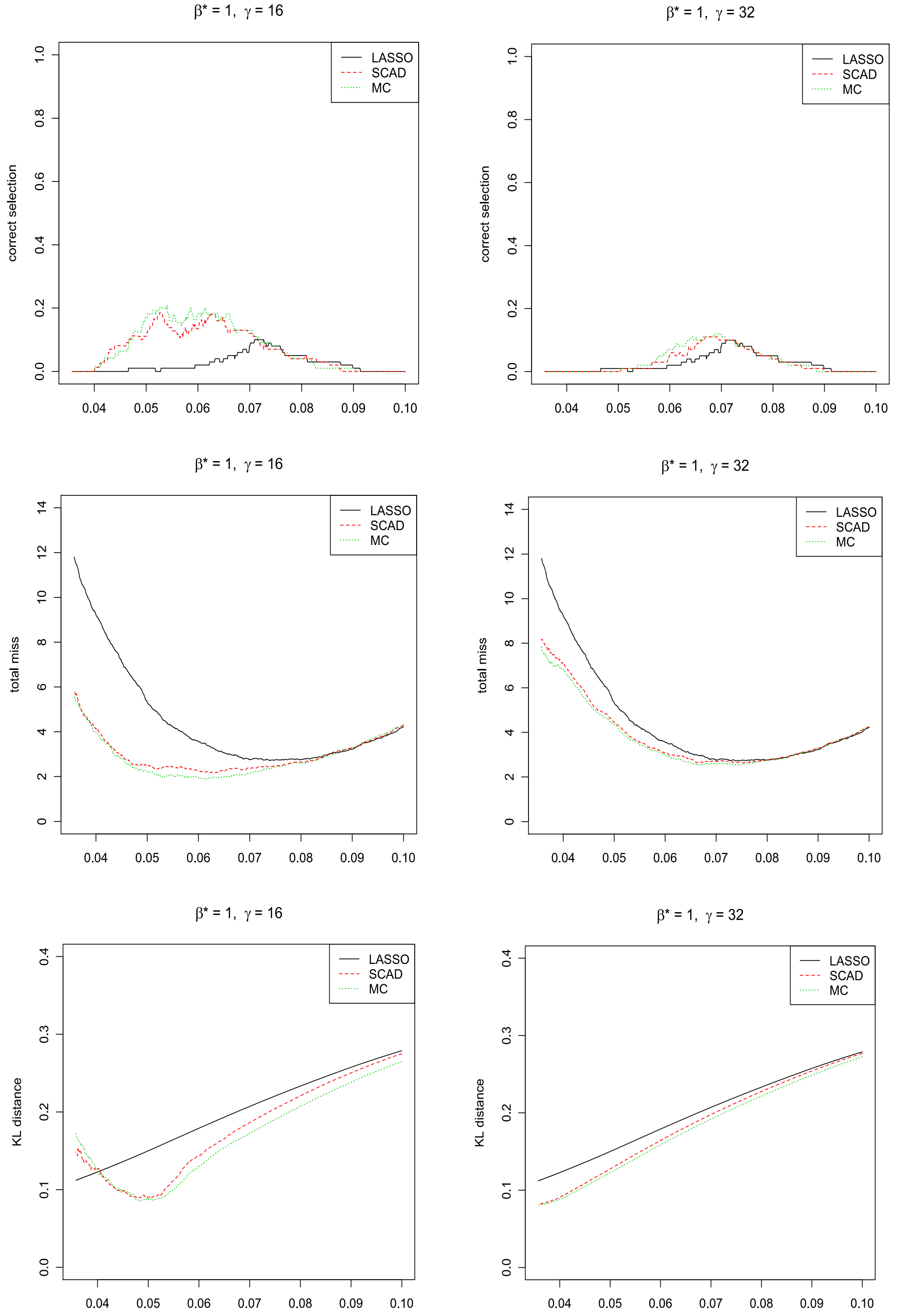

Figure 4.4: The comparison between the LASSO (black, solid), SCAD (red, dashed) and MC (green, dotted) in logistic regression based on 100 replications. The correct selection probability $\overline{CS}$, total miss $\overline{TM}$ and Kullback-Leibler divergence $\overline{KL}$ are plotted against the penalty level λ. Top panels: $\overline{CS}$; middle panels: $\overline{TM}$; bottom panels: $\overline{KL}$; left panels: $\gamma = 16$; right panels: $\gamma = 32$. Parameters: $n = 200$, $p = 100$, $d^o = 10$, $\beta^* = 1$ and $\Delta = 0.005$.

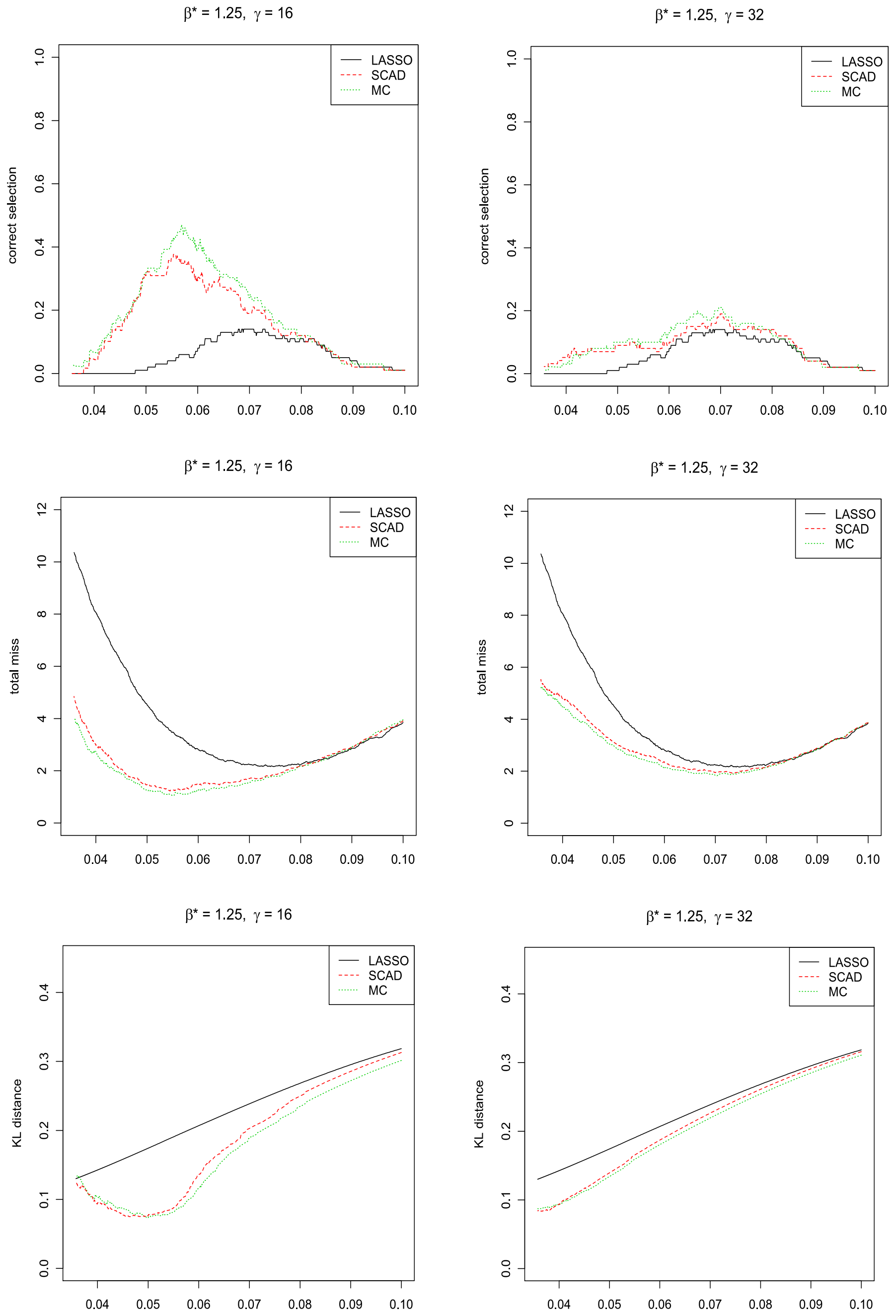

Figure 4.5: The comparison between the LASSO (black, solid), SCAD (red, dashed) and MC (green, dotted) in logistic regression based on 100 replications. The correct selection probability $\overline{CS}$, total miss $\overline{TM}$ and Kullback-Leibler divergence $\overline{KL}$ are plotted against the penalty level λ. Top panels: $\overline{CS}$; middle panels: $\overline{TM}$; bottom panels: $\overline{KL}$; left panels: $\gamma = 16$; right panels: $\gamma = 32$. Parameters: $n = 200$, $p = 100$, $d^o = 10$, $\beta^* = 1.25$ and $\Delta = 0.005$.

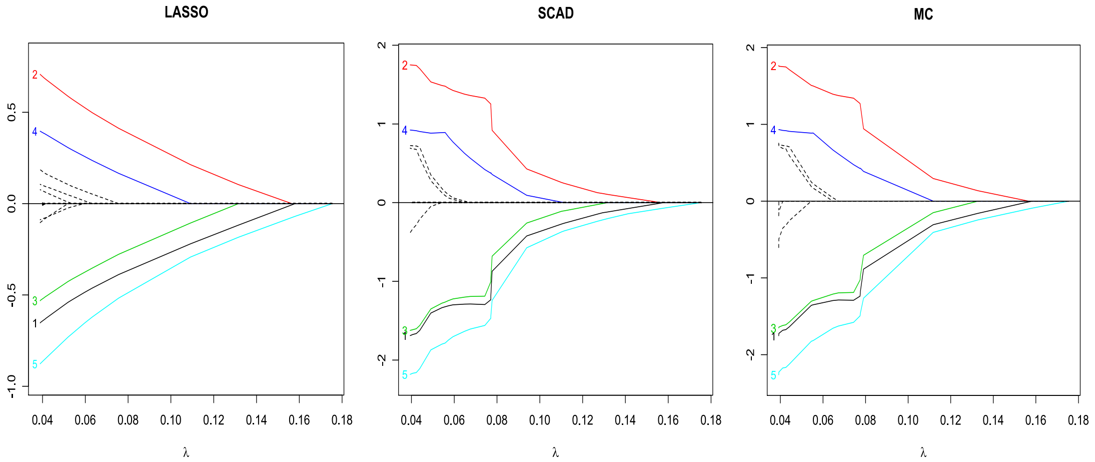

Figure 4.6: The solution paths of the LASSO, SCAD and MC with one replication for $n = 200$, $p = 100$, $d^o = 5$, $\beta^* = 1.25$, $\gamma = 16$ and $\Delta = 0.005$. The estimates $\widehat{\beta}_j$ are plotted against the penalty level λ. The colored solid curves correspond to the covariates with $\beta_j \neq 0$. The dashed curves correspond to the covariates which do not influence the response with $\beta_j = 0$. The solutions plotted are build up in 4000 steps. The SCAD and the MC paths are almost indistinguishable with each other, while the LASSO paths are different from the SCAD and MC.

the estimated values of coefficients of true variables roughly "keep" after several noisy variables are incorporated, while the LASSO paths keep increasing.

An explanation for setting $\gamma = 16$ is as follows: in logistic regression, the variance $p_i(1-p_i)$ is no more than 1/4. Therefore, the eigenvalues of $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}/n$ are approximately 4 more times smaller than those of $\boldsymbol{X}^T\boldsymbol{X}/n$. To roughly keep the convexity as in the linear model, the value of γ in logistic model should be more than 4 times of 3.7, the typical value of γ suggested for linear model.

4.4.3 S&P 500 Index data

The S&P 500 (Standard & Poor's 500) is a market-value-weighted index of 500 stocks that are traded on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and the NASDAQ National Market System. Companies selected for the S&P 500 Index (SPX) are representative of important industries within the U.S. economy and many also are the leaders of their industries. The

SPX is the summation of the weighted stocks prices,

$$I(s) = \sum_{j=1}^{500} w_j X^j(s), \tag{4.32}$$

where $I(s)$ is the SPX at time s, $X^j(s)$ is the price of the j-th stock at time s, and w_j are the weights which make each company's influence on the SPX performance directly proportional to their market values. In evaluations and performance charts of stocks and mutual funds, the SPX is regarded as one important baseline for comparison. For example, a performance chart of a mutual fund will show the SPX along with their financial product. Many ETF's (exchange-traded funds) attempt to replicate the performance of the SPX. However, it would cost too much effort and capital to hold all S&P stocks to replicate the SPX. Thus, one way to mimic the SPX is to hold a subset of all stocks and figure out their weights simultaneously.

In this experiment, we select a subset of all S&P stocks and figure out a linear combination of them to estimate the future index. We collect close index and close prices of S&P stocks from July 26, 2007 to July 25, 2008. We use a moving window method to compare the one-step replication performances of the LASSO, SCAD and MC. In detail, denote $\boldsymbol{Z}(s) = (I(s), X^j(s), j = 1, \ldots, p)$, $s = 1, \ldots, n$ as the raw close-of-day data of the s-th trading day. Let $(\boldsymbol{Z}(s), s = s_0, \ldots, s_0+m-1)$ be the raw data of m consecutive days. We fit penalized linear models to estimate the regression coefficients $\widehat{\boldsymbol{w}}(d) = (\widehat{w}_j(d), j = 1, \ldots, p)$ where d denotes the number of nonzero $\widehat{w}_j$. Thus, a subset of d stocks is selected. We next compute the one-step replication error $e_m(s_0 + m, d)$ on next data $\boldsymbol{Z}(s_0 + m)$,

$$e_m(s_0 + m, d) \equiv \Big| I(s_0 + m) - \sum_{j=1}^{p} \widehat{w}_j(d) X^j(s_0 + m) \Big|.$$

In Table 4.4, we report the average replication errors $\bar{e}_m(d) \equiv \sum_{s_0=1}^{n-m} e_m(s_0 + m, d)/(n-m)$ with $n = 253$, $m = 200$ and $d = 5k, k = 2, \ldots, 13$. From Table 4.4, it can be seen that with almost all $d \le 65$, the average error of either the SCAD or MC improves over the LASSO. An interesting phenomenon is that when the sizes

Table 4.4: One-step replication performances of the LASSO, SCAD and MC with the moving window method. Average replication errors $\bar{e}_m(d) \equiv \sum_{s_0=1}^{n-m} e_m(s_0 + m, d)/(n - m)$ with the window size $m = 200$ and stocks subset of sizes $d = 5k$, $k = 2, \ldots, 13$ are reported.

d	10	15	20	25	30	35
LASSO	16.08	14.40	13.44	11.41	8.74	5.90
SCAD	16.08	14.40	13.44	11.41	8.74	5.90
MC	16.06	14.42	13.07	10.75	7.73	4.79
d	40	45	50	55	60	65
LASSO	4.77	3.91	3.51	3.28	2.96	2.77
SCAD	4.77	3.91	3.52	3.38	3.18	2.92
MC	4.08	3.73	3.48	3.04	2.99	2.68

of the stock subsets are small, the replication results of the LASSO and SCAD are same. This is not surprising since the penalty functions $\rho(t)$ of the LASSO and SCAD are identical when $0 \leq t \leq 1$ (see Figure 4.1).

4.4.4 South African heart disease data

As another real data example, we consider the South African heart disease data as used in [42]. In this dataset, there are $p = 9$ variables and $n = 462$ observations. The response is a binary variable which indicates the presence ($y = 1$) or absence ($y = 0$) of myocardial infarction.

Figure 4.7 shows the approximated paths of the three methods. The ℓ_1 norm of the coefficients forms the x-axis and the fitted coefficients $\widehat{\beta}$ are plotted against the ℓ_1 norm. The data are processed so that each feature has mean zero and ℓ_2 norm $\sqrt{n}$. In each panel, the solutions are computed in 30000 steps (with the step size $\Delta = 0.01$ and $\gamma = 16$). The SCAD paths and MC paths are almost identical.

In the second part of this experiment, we add 100 noisy variables $\boldsymbol{x}^j, j = 10, \ldots, 109$ to the original 9 variables $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^9$, so that the design is $\boldsymbol{X} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^{109})$. To generate the noisy variables $\boldsymbol{X}^{no} = (\boldsymbol{x}^{10}, \ldots, \boldsymbol{x}^{109})$, we first generate an $n \times p^*$ pool random matrix $\boldsymbol{X}^{pl}$ with each cell iid standard normal random variable. In this experiment, $(n, p^*) = (462, 500)$. Still, $\boldsymbol{X}^{pl}$ is

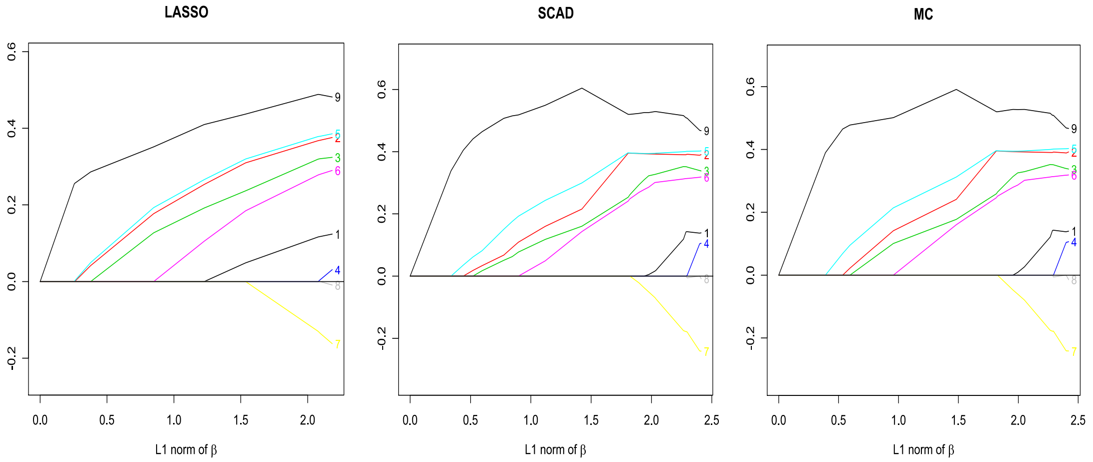

Figure 4.7: The solutions paths of the various penalized logistic regression models for the South African heart disease data. The fitted coefficients are plotted against the ℓ_1 norm. Still, the SCAD and MC paths are almost identical.

standardized so that each column has mean zero and ℓ_2 norm $\sqrt{n}$. The noisy variables $\boldsymbol{X}^{no}$ is then generated by greedy sequential group sampling from $\boldsymbol{X}^{pl}$. Each sampling group consists of 10 most correlated vectors from the remaining columns of $\boldsymbol{X}^{pl}$. For the m-th group, we sample from the remaining $510 - 10m$ columns one member as $\boldsymbol{x}^{10m-9}$ and 9 more to maximize the absolute correlation $|\boldsymbol{x}^{jT}\boldsymbol{x}^{10m-9}|/n, j = 10m-8, \ldots, 10m, m = 1, \ldots, 10$.

We apply logistic regression model with the ℓ_1, the SCAD and MC methods to the heart disease data with noisy variables and exhibit their model selection properties in Table 4.5. Our parameter here is again $(\gamma, \Delta) = (16, 0.01)$. When the selection is only among the original variables, the three methods select all 9 variables in 25000 steps. With the presence of noisy variables, all three procedures select variables 2, 3, 5 and 9 in the very beginning stage without the influence of the noises. The LASSO selects variable 1 somewhat later and variables 4, 7 and 8 after lots of noisy variables, while the SCAD and MC select variable 7 somewhat later and variables 1, 4 and 8 much later. These observations may imply that variables 2, 3, 5 and 9 are important variables to explain their joint effect on the prevalence of myocardial infarction. This selection properties are the same as the result of stepwise logistic regression fit summarized in Table 4.3 in [42].

Table 4.5: The selection order for the original 9 variables. The “w.n.” indicates the selection with noises: 100 simulated noisy variables are added to the original 9 variables

Variables	1	2	3	4	5	6	7	8	9
LASSO	6	3	4	8	2	5	7	9	1
LASSO (w.n.)	21	3	4	90	2	6	42	60	1
SCAD	7	3	4	9	2	5	6	8	1
SCAD (w.n.)	46	3	4	88	2	5	27	49	1
MC	7	3	4	9	2	5	6	8	1
MC (w.n.)	45	3	4	87	2	5	26	48	1

Table 4.6: The selection order for the original 9 variables. The “w.n.” indicates the selection with noises: 100 simulated noisy variables are added to the original 9 variables. Artificial responses y_i are generated from bernoulli distributions with probabilities $P(y_i = 1|\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}})/(1+\exp(\boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}}))$ where $\widetilde{\beta}_j = \widehat{\beta}_j^{MLE}, j = 1,\ldots,9$ and $\widetilde{\beta}_j = 0, j = 10,\ldots,109$

Variables	1	2	3	4	5	6	7	8	9
LASSO	5	4	3	9	2	6	7	8	1
LASSO (w.n.)	5	4	3	75	2	7	17		1
SCAD	6	7	3	9	2	4	5	8	1
SCAD (w.n.)	12	16	3	42	2	4	9		1
MC	6	7	3	9	2	4	5	8	1
MC (w.n.)	11	15	3	42	2	4	8		1

In the third part of this experiment, we generate artificial responses y_i from bernoulli variables whose probabilities of success are $P(y_i = 1|\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}})/(1+\exp(\boldsymbol{x}_i^T\widetilde{\boldsymbol{\beta}}))$ where $\widetilde{\beta}_j = \widehat{\beta}_j^{MLE}, j = 1,\ldots,9$ and $\widetilde{\beta}_j = 0, j = 10,\ldots,109$. The design matrix $\boldsymbol{X} = (\boldsymbol{x}^1,\ldots,\boldsymbol{x}^{109})$ is the same as that used in the second part of this experiment. We run the same six procedures as listed in Table 4.6 to the artificial data. We set $(\gamma,\Delta) = (16, 0.01)$. In the first 25000 steps, with the artificial data and the presence of noisy variables, all the three procedures dismiss variable 8. The LASSO select variables 1, 2, 3, 5 and 9 in the very beginning stage without the influence of the noises, while the SCAD and MC achieve so for variables 3, 5, 6 and 9. The LASSO selects variable 7 somewhat later and variables 4 very late, while the SCAD and MC select variable 1 and 2 somewhat later and variables 4 very late.

4.5 Discussion

In this chapter, we present the GPLUS algorithm to compute the solution paths of the concave-penalized negative log-likelihood. Generally speaking, the GPLUS algorithm may be applied to any situation where the objective function $\psi(\boldsymbol{\beta})$, whether is likelihood or not, has continuous first two derivatives.

As pointed by [86], the ℓ_1 and concave penalties represent the two main streams of penalization method for variable selection in the recent literature. The ℓ_1 penalty results in convex minimization problem thus it is computational more friendly, while the concave method is asymptotically unbiased and enjoys the oracle properties. Applying the GPLUS algorithm to these penalty-based variable selection methods, the MC and SCAD generate much better variable selection accuracy than the LASSO in sparse linear models. In sparse binary logistic regression models, concave penalization approach still shows considerable improvement over the ℓ_1 method. The surprising aspect is that in section 4.4.1, with proper choice of parameters of controlling the convexity of penalized least squares, the MC shows great improvement over the SCAD. In logistic regression, such improvement is much less prominent because of the loss of convexity.

We note that the "one-at-a-time" condition $|C_1^{(k)}| = 1$ holds almost everywhere. That means the boundary crossing never involves more than a single index j with $\eta_j^{(k)} \neq 0$. Since one-at-a-time condition, perhaps with some jitter of Δ, holds to all practical situations, we do not consider the many-at-a-time problems in the GPLUS procedure. Instead, even if the one-at-a-time condition does not hold, we admit that the crossings happen for all the critical indices in $C_1^{(k)}$.

The simulation and theoretical results in [78] show that the universal penalty level $\lambda = \widehat{\sigma}\sqrt{2(\log p)/n}$ is nearly the optimal choice for variable selection in linear model (4.28) with $N(0, \sigma^2)$ errors. When $p < n$, the mean residual squares $\|\boldsymbol{y} - \widetilde{\boldsymbol{u}}\|^2/\{n - \text{rank}(\boldsymbol{X})\}$ provides a good estimator of σ^2 where $\widetilde{\boldsymbol{u}}$ is the projection of $\boldsymbol{y}$ to the linear span of the design vectors $\{\boldsymbol{x}^j, j \leq p\}$. Similarly, in logistic regression, the optimal penalty level is

$$\lambda^* = \Big(\sum_{i=1}^{n} p_i(1-p_i)/n \Big)^{1/2} \sqrt{2(\log p)/n}, \tag{4.33}$$

where p_i is the probability of success in (4.29). In our investigation, estimating λ^* based on the MLE of $\boldsymbol{\beta}$ does not work well since MLE of logistic regression typically generates poor estimation even with moderately large p. We have also tried to carry out estimation of $\boldsymbol{\beta}$ and variable selection simultaneously but gained limited improvements in selection accuracy. We believe that a good estimation of λ^* should based on the direct estimation of the first factor on the right hand side of (4.33) instead of $\boldsymbol{\beta}$. With proper choice of penalty level and some general regularity conditions, we expect that the asymptotic error bounds for variable selection can be established. In this chapter we focus on the algorithm without the discussion of theories.

4.6 Proof

Proof of Theorem 1. To state the proof, we will use some more explicit notation. Denote $\widehat{\boldsymbol{b}}(\tau^{(j)}) \equiv \boldsymbol{b}^{(j)}$ and $\boldsymbol{s}(\widehat{\boldsymbol{b}}(\tau^{(j)}), \tau^{(j)}) \equiv \boldsymbol{s}^{(j)}$ where $\boldsymbol{b}^{(j)}$ is the j-th turning point computed by the GPLUS algorithm as in Section 3.2. For simplicity, denote $\rho_j \equiv \|\widehat{\boldsymbol{b}}(\tau^{(j)}) - \boldsymbol{b}(\tau^{(j)})\|$.

By Taylor expansion and the GPLUS algorithm

$$\begin{aligned} \boldsymbol{b}(\tau^{(k_0+1)}) &= \boldsymbol{b}(\tau^{(k_0)}) + (\tau^{(k_0+1)} - \tau^{(k_0)})\boldsymbol{s}(\boldsymbol{b}(\tau^{(k_0)}), \tau^{(k_0)}) \\ &\quad + \frac{1}{2}\Big(\frac{d}{d\tau}\boldsymbol{s}(\boldsymbol{b}(\tau), \tau)|_{\tau=\widetilde{\tau}}\Big)(\tau^{(k_0+1)} - \tau^{(k_0)})^2, \qquad (4.34) \\ \widehat{\boldsymbol{b}}(\tau^{(k_0+1)}) &= \widehat{\boldsymbol{b}}(\tau^{(k_0)}) + (\tau^{(k_0+1)} - \tau^{(k_0)})\boldsymbol{s}(\widehat{\boldsymbol{b}}(\tau^{(k_0)}), \tau^{(k_0)}). \qquad (4.35) \end{aligned}$$

where $\widetilde{\tau} = \tau^{(k_0)} + \theta(\tau^{(k_0+1)} - \tau^{(k_0)})$ with some $0 < \theta < 1$. We choose Δ small enough such that $\rho_{k_0} \le \delta$ and $(\rho_{k_0} + M_2M_3\Delta)\exp(M_1M_3) \le \delta$. Comparing (4.34) with (4.35) and utilizing the conditions (ii), (iii) and (iv), we have

$$\begin{aligned}
\rho_{k_0+1} &\leq \rho_{k_0} + \Delta^{(k_0+1)}\|\boldsymbol{s}(\widehat{\boldsymbol{b}}(\tau^{(k_0+1)}), \tau^{(k_0+1)}) - \boldsymbol{s}(\boldsymbol{b}(\tau^{(k_0)}), \tau^{(k_0)})\| + M_2(\Delta^{(k_0+1)})^2 \\
&\leq (1 + \Delta^{(k_0+1)}M_1)\rho_{k_0} + M_2(\Delta^{(k_0+1)})^2 \\
&\leq (\rho_{k_0} + M_2(\Delta^{(k_0+1)})^2)\exp(M_1\Delta^{(k_0+1)}) \\
&\leq (\rho_{k_0} + M_2M_3\Delta)\exp(M_1M_3) \leq \delta. \qquad (4.36)
\end{aligned}$$

Generally, when $\rho_j \leq \delta$ for $j = k_0+1, \ldots, k-1$, by induction we have

$$\begin{aligned}
\rho_k &\leq M_2(\Delta^{(k)})^2 + \rho_{k-1}(1 + \Delta^{(k)}M_1) \\
&\leq M_2(\Delta^{(k)})^2 + M_2(\Delta^{(k-1)})^2(1 + \Delta^{(k)}M_1) + \rho_{k-2}(1 + \Delta^{(k-1)}M_1)(1 + \Delta^{(k)}M_1) \\
&\leq M_2(\Delta^{(k)})^2 + M_2(\Delta^{(k-1)})^2(1 + \Delta^{(k)}M_1) \\
&\quad + M_2(\Delta^{(k-2)})^2(1 + \Delta^{(k-1)}M_1)(1 + \Delta^{(k)}M_1) \\
&\quad + \cdots + M_2(\Delta^{(k_0+1)})^2 \prod_{j=k_0+2}^{k}(1 + \Delta^{(j)}M_1) + \rho_{k_0}\prod_{j=k_0+1}^{k}(1 + \Delta^{(j)}M_1) \\
&\leq (\rho_{k_0} + M_2\sum_{j=k_0+1}^{k}(\Delta^{(j)})^2)\exp(M_1\sum_{j=k_0+1}^{k}\Delta^{(j)}) \\
&\leq (\rho_{k_0} + M_2M_3\Delta)\exp(M_1M_3) \leq \delta,
\end{aligned}$$

Hence, we have proved that in the block $\boldsymbol{\eta}$, when Δ is small enough, the fact that all the previous estimated turning points $\widehat{\boldsymbol{b}}(\tau^{(j)}), j = k_0, \ldots, k-1$ are located in the δ-"tube" around the true paths $\boldsymbol{b}(\tau^{(j)})$ will result in the next turning point $\widehat{\boldsymbol{b}}(\tau^{(k)})$ living in the δ-tube around $\boldsymbol{b}(\tau^{(k)})$. Thus, our induction can move on. Notice that $(\rho_{k_0} + M_2M_3\Delta)\exp(M_1M_3) \to 0$ as $\Delta \to 0$, which implies that $\rho_k \to 0$ as $\Delta \to 0$. This completes the proof. □

References

[1] Abramovich, F., Benjamini, Y., Donoho, D.L. and Johnstone, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584-653.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd International Symposium on Information Theory,* V. Petrov and F. Csáki, eds. 267-281. Akadmiai Kiadó, Budapest.

[3] Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical inference under order restrictions; the theory and application of isotonic regression.* Wiley, New York.

[4] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B.* **57** 289-300.

[5] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3** 203-268.

[6] Borell, C. (1975). The Brunn-Minkowski inequality in Gaussian space. *Invent. Math.* **30** 207-216.

[7] Brandwein, A.C. and Strawderman, W.E. (1990). Stein Estimation: The Spherically Symmetric Case. *Statist. Science* **5** 356-369.

[8] Brideau, C., Gunter, B., Pikounis, B. and Liaw, A. (2003). Improved statistical methods for hit selection in high-throughput screening. *J. Biomolecular Screening* **8** 634-647.

[9] Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855-903.

[10] Brown, L.D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Appl. Statist.* **2** 113-152.

[11] Brown, L.D. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685-1704.

[12] Cai, T.T. (1999). Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898-924.

[13] Cai, T.T. (2002). On block thresholding in wavelet regression. *Statist. Sinica* **12** 1241-1273.

[14] Cai, T.T. and Silverman, B.W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhyā Ser. B* **63** 127-148.

[15] Cai, T.T. and Zhou, H.H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.* **37** 569-595

[16] Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313-2351.

[17] Carathéodory, C. (1911). Über den Variabilitätsbereich der Fourierschen Konstanten von positiven harmonischen Funktionen. *Rend. Circ. Mat. Palermo* **32** 193-217.

[18] Cover, T.M. (1984). An algorithm for maximizing expected log investment return. *IEEE Trans. Inform. Theory* **30** 369-373.

[19] Daubechies, I. (1992). *Ten Lectures on Wavelets.* SIAM, Philadelphia, PA.

[20] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B* **39** 1-38.

[21] Donoho, D.L. and Johnstone, I.M. (1994a). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99** 277-303.

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

[23] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200-1224.

[24] Donoho, D.L. and Johnstone, I.M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879-921.

[25] Donoho, D.L., Johnstone, I.M., Hoch, J.C. and Stern, A.S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41-81.

[26] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301-369.

[27] Efroimovich, S. YU. and Pinsker, M.S. (1984). A learning algorithm for nonparametric filtering. *Autom. Remote Control* **11** 58-65.

[28] Efron, B. (2003). Robbins, empirical Bayes and microarrays. *Ann. Statist.* **31** 366-378.

[29] Efron, B., Hastie, T., Johnstone, I.M. and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32** 407-499.

[30] Efron, B. and Morris, C.N. (1972). Empirical Bayes on vector observations: An extension of Steins method. *Biometrika* **59** 335-347.

[31] Efron, B. and Morris, C.N. (1973). Steins estimation rule and its competitorsCan empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117-130.

[32] Efron, B. and Tibshirani, R. (2002) Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology* **23** 70-86.

[33] Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001) Empirical bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151C1160.

[34] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348-1360.

[35] Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32** 928-961.

[36] Foster, D.P. and George, E.I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947-1975.

[37] Genkin, A., Lewis. D.D. and Madigan, D. (2004). Large-scale Bayesian logistic regression for text categorization. Preprint.

[38] George, E.I. (1986). Mimimax multiple shrinkage estimation. *Ann. Statist.* **14** 288-305.

[39] Ghosal, S. and van der Vaart, A.W. (2001). Entropies and rate of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233-1263.

[40] Ghosal, S. and van der Vaart, A.W. (2007). Posterior convergence rates for Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697-723.

[41] Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality: The Third Erich L. Lehmann Symposium,* J. Rojo, Ed., Institute of Mathematical Statistics, Lecture Notes-Monograph Series **57** 266-275.

[42] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning: data mining, inference, and prediction.* Springer, New York.

[43] Hunter, D.R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617-1642.

[44] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361-379. Univ. of California Press, Berkeley.

[45] Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647-1684.

[46] Johnstone, I.M. (1994). Minimax Bayes, asymptotic minimax and sparse wavelet priors. In *Statistical Decision Theory and Related Topics V* (S. Gupta and J. Berger, eds.) 303-326. Springer, New York.

[47] Johnstone, I.M. and Silverman, B.W. (2004). Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594-1649.

[48] Johnstone, I.M. and Silverman, B.W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700-1752.

[49] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887-906.

[50] Lokhorst, J. (1999). The lasso and generalised linear models. *Technical report*, University of Adelaide.

[51] Meishausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[52] Meishausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246-270.

[53] Meyer, Y. (1992). *Wavelets and Operators.* Cambridge Univ. Press.

[54] Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47-55.

[55] Osborne, M., Presnell, B. and Turlach, B. (2000a). A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20** 389-403.

[56] Osborne, M., Presnell, B. and Turlach, B. (2000b). On the LASSO and its dual. *J. Comput. Graph. Statist.* **9** 319-337.

[57] Park, M. and Hastie, T. (2007). An L1 regularization-path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69** 659-677.

[58] Pinsker, M.S. (1980). Optimal Filtration of square-integrable signals in Gaussian White Noise. *Problems of Information Transmission* **16** 120-133.

[59] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. *Proc. Second Berkeley Symp. Math. Statist. Probab.* **1** 131-148. Univ. of California Press, Berkeley.

[60] Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157-163. Univ. of California Press, Berkeley.

[61] Robbins, H. (1964). The empirical Bayes approach to statistical decision problem. *Ann. Math. Statist.* **35** 1-20.

[62] Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713-723.

[63] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35** 1012-1030.

[64] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157-163. Univ. of California Press, Berkeley.

[65] Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385-388.

[66] Tang, W. and Zhang, C.-H. (2005). Bayes and empirical Bayes approaches to controlling the false discovery rate. *Technical Report 2005-004*, Department of Statistics and Biostatistics, Rutgers University.

[67] Tang, W. and Zhang, C.-H. (2007). Empirical Bayes methods for controlling the false discovery rate with dependent data. In *Complex Datasets and Inverse Problems: Tomography, Networks, and Beyond*, R. Liu, W. Strawderman and C.-H. Zhang, Eds., Institute of Mathematical Statistics, Lecture Notes-Monograph Series **54** 151-160.

[68] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58** 267-288.

[69] Triebel, H. (1992). *Theory of Function Spaces. II.* Birkhäuser, Basel.

[70] van der Vaart, A.W. and Wellner, J.A. (1996). Weak Convergence and Empirical Processes. Springer, New York.

[71] Vardi, Y. and Lee, D. (1993). From image deblurring to optimal investment: maximum likelihood solutions for positive linear inverse problem (with discussion). *J. Roy. Statist. Soc. B* **55** 569-612.

[72] Wasserman, L. (2007). *All of nonparametric statistics.* Springer, New York.

[73] Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. *Ann. Statist.* **18** 806-831.

[74] Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* **7** 181-193.

[75] Zhang, C.-H. (2003). Compound decision theory and empirical Bayes method. *Ann. Statist.* **33** 379-390.

[76] Zhang, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33** 54-100.

[77] Zhang, C.-H. (2007a). Continuous generalized gradient descent. *J. Comput. Graph. Statist.* **16** 761-781.

[78] Zhang, C.-H. (2007b). Penalized linear unbiased selection. *Technical Report 2007-003*, Department of Statistics, Rutgers University.

[79] Zhang, C.-H. (2008a). Discussion of "one-step sparse estimates in nonconcave penalized likelihood models". *Ann. Statist.* **36** 1553-1560.

[80] Zhang, C.-H. (2008b). Generalized maximum likelihood estimation of normal mixture densities. *Statist. Sinica*, to appear.

[81] Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567-1594.

[82] Zhao, P., Rocha, G.V. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* To appear.

[83] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Research* **7** 2541-2567.

[84] Zhao, P. and Yu, B. (2007). Stagewise Lasso. *J. Machine Learning Research*, **8** 2701-2726.

[85] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418-1429.

[86] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509-1533.

Vita

Wenhua Jiang

2000 Graduated from Fudan University High School, Shanghai, China.

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

2009 Ph.D. in Statistics, Rutgers, The State University of New Jersey, New Brunswick, New Jersey.