Topics in Penalized Estimation

by Elizabeth Danielle Schifano

TOPICS IN PENALIZED ESTIMATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Elizabeth Danielle Schifano

August 2010

TOPICS IN PENALIZED ESTIMATION

Elizabeth Danielle Schifano, Ph.D.

Cornell University 2010

The use of regularization, or penalization, has become increasingly common in high-dimensional statistical analysis over the past several years, where a common goal is to simultaneously select important variables and estimate their effects. This goal can be achieved by minimizing some parameter-dependent "goodness of fit" function (e.g., negative loglikelihood) subject to a penalization that promotes sparsity. Penalty functions that are nonsmooth (i.e., not differentiable) at the origin have received substantial attention, arguably beginning with LASSO (Tibshirani, 1996).

This dissertation consists of three parts, each related to penalized estimation. First, a general class of algorithms is proposed for optimizing an extensive variety of nonsmoothly penalized objective functions that satisfy certain regularity conditions. The proposed framework utilizes the majorization-minimization (MM) algorithm as its core optimization engine. The resulting algorithms rely on iterated soft-thresholding, implemented componentwise, allowing for fast, stable updating that avoids the need for any high-dimensional matrix inversion. Local convergence theory is established for this class of algorithms under weaker assumptions than previously considered in the statistical literature. The second portion of this work extends the MM framework to finite mixture regression models, allowing for penalization among the regression coefficients within a potentially unknown number of components. Finally, a hierarchical structure imposed on the penalty parameter provides new motivation for the Minimax Concave Penalty of Zhang (2010). Frequentist and Bayesian risk of the MCP thresholding estimator and several other thresholding estimators are compared and explored in detail.

BIOGRAPHICAL SKETCH

Elizabeth Danielle Schifano, native to Buffalo, NY, completed her undergraduate studies at Cornell University. She received her Bachelor of Science, Summa Cum Laude, from the Department of Biological Statistics and Computational Biology (formerly known as the Department of Biometry and Statistics) within the College of Agricultural and Life Sciences.

Elizabeth also received her Master of Science degree in Statistics from Cornell University. Upon completion of her doctoral degree, she ultimately hopes to pursue a career applying her statistical knowledge to the biological, particularly genomics, arena.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Professor Robert L. Strawderman, for his counsel, help and encouragement throughout my years at Cornell. He has been a tremendous advocate, both personally and professionally, and has been an inspiring influence in my work. I am grateful to Dr. Strawderman for his patience and for the countless hours spent helping me work through my questions. He is a brilliant, dedicated mentor and I am most fortunate to have had the opportunity to learn from him.

I am also indebted to my dissertation committee members, Drs. Martin T. Wells and James G. Booth, for their invaluable guidance and intuition in this work and beyond. It was a great pleasure and honor to work with such distinguished faculty, and I have truly grown and matured as a statistician as a result.

I am grateful to my family for their many thoughts and prayers, especially within the last year. My parents, Louis P. Schifano, Mary P. Andreozzi, and Phillip N. Andreozzi, Jr., deserve special mention for their endless support in all of my endeavors, both academic and otherwise. Their persistent confidence in me made the impossible seem possible; for this, I can not thank them enough.

Finally, words alone cannot express my appreciation and thanks to my fiancé, Joshua P. Egnatz. For his support, encouragement, companionship, and for just being, he has my everlasting love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Variable selection is an important and challenging issue in the rapidly growing realm of high-dimensional statistical modeling. In such cases, it is often of interest to identify a few important variables in a veritable sea of noise. Modern methods, increasingly based on the principle of penalized likelihood estimation applied to high dimensional regression problems, attempt to achieve this goal through an adaptive variable selection process that simultaneously permits estimation of regression effects. Indeed, the literature on the penalization of a "goodness of fit" function (e.g., negative loglikelihood), with a penalty singular at the origin, is quickly becoming vast, proliferating in part due to the consideration of specific combinations of data fidelity (i.e., goodness-of-fit) and penalty functions, the associated statistical properties of resulting estimators, and the development of several combination-specific optimization algorithms, (e.g., Tibshirani, 1996; Fan and Li, 2001; Zou and Hastie, 2005; Zou, 2006; Park and Hastie, 2007; Friedman et al., 2008; Zou and Zhang, 2009).

With this in mind, a unified optimization framework is proposed that appeals to the Majorization-Minimization (MM) algorithm (Lange et al., 2000) as the primary optimization tool. The resulting class of algorithms is referred to as MIST, an acronym for Minimization by Iterative Soft Thresholding. The MM algorithm has been considered before for solving specific classes of singularly penalized likelihood estimation problems (e.g., Daubechies et al., 2004; Hunter and Li, 2005; Zou and Li, 2008); to a large extent, this work is motivated by these ideas. A distinct advantage of the proposed work is the exceptional versatility of the class of MIST algorithms, their associated ease of implementation and numerical stability, and the development of a fixed point convergence theory that permits weaker assumptions than existing papers in this area.

The MIST algorithm can also be used in finite mixture regression (FMR) in conjunction with the Expectation Maximization (EM) algorithm as a way to determine mixture component membership. As the EM algorithm can be viewed as a special case of the MM algorithm, the FMR problem fits naturally within the MM theory and framework, with a few minor complicating factors. Selection of the number of components is a common problem faced in the FMR literature, but an iterative procedure is proposed to penalize the regression coefficients within each component, for a potentially unknown number of components.

Critically important to any penalized optimization problem is the selection of appropriate tuning or penalty parameters. Typically, these are assumed to be fixed and known throughout the optimization process. In practice, these parameters are often adaptively chosen a posteriori by minimizing some criteria using such techniques as AIC, BIC, C_p, k-fold cross-validation, generalized cross-validation, etc. An alternative idea, explored recently in Park and Casella (2008) and Strawderman and Wells (2010), is to take a more Bayesian approach and equip the penalty parameter with a prior distribution. In such cases, the issue of tuning parameter selection shifts to the issue of prior hyperparameter selection. Interestingly, an appropriate choice of prior on the L_1 penalty parameter can lead to the Minimax Concave Penalty (MCP) thresholding estimator of Zhang (2010), with hyperparameters corresponding the the traditional tuning parameters.

The dissertation is structured as follows. Chapter 2 contains relevant background information that will, in part, serve as a reference for later chapters. In Chapter 3, a general MM local convergence theory is presented for objective functions that are 'non-smooth' in a certain sense. Chapter 4 is devoted to MIST and includes sufficient conditions for local convergence of the MM algorithm (as specified in Chapter 3) applied to a large class of data-fidelity and non-smooth penalty functions; specialized versions

of this general algorithm, demonstrating in particular how the minimization step of the MM algorithm can be carried out using iterated soft-thresholding; detailed simulation results; and an application in survival analysis to Diffuse Large B Cell Lymphoma expression data (Rosenwald et al., 2002). In Chapter 5, the MIST algorithm is extended for use in FMR models. A general algorithm is presented, along with a modification for mixtures of linear regression models with common unknown variance. Simulation results and analysis of the popular ozone dataset of Breiman and Friedman (1985) are included to demonstrate applicability. In the spirit of Park and Casella (2008) and building on the work of Strawderman and Wells (2010), Chapter 6 explores a hierarchical model motivation to the univariate Minimax Concave Penalty (MCP, Zhang, 2010) thresholding estimator, with a detailed risk assessment and simulation study for MCP and other univariate thresholding estimators. A concluding discussion is provided in Chapter 7.

CHAPTER 2

USEFUL BACKGROUND INFORMATION

The following sections will provide a helpful review and additional information that will serve as a useful reference for later chapters. Section 2.1 is relevant to the entire dissertation, as it provides a brief overview of variable selection using penalized likelihoods. Specific examples of penalized likelihoods will be explored particularly in Chapter 4. Sections 2.1.1 and 2.1.2 are included to help tie the three parts of this work together. Section 2.2 reviews the basic principle of the Majorization-Minimization (MM) algorithm, upon which Chapters 3, 4, and 5 heavily rely. Finally, Section 2.3 collects relevant facts and properties for the Clarke subdifferential theory used in Chapters 3, 4, and 5.

2.1 Variable Selection and Penalized Likelihoods

Consider the linear regression model

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

where $\mathbf{y}$ is an $N \times 1$ vector of responses, $\mathbf{X}$ is an $N \times p$ matrix of covariates, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown coefficients. It is assumed that $\boldsymbol{\epsilon}$ is an $N \times 1$ vector of independent elements $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, N$. When the number of observations N is greater than the number of covariates p, the ordinary least-squares estimate is given by $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Over the years, many strategies have been developed to weed out unnecessary covariates. Such strategies often involve optimizing a selection criterion, such as AIC, BIC, Mallow's C_p, PRESS, adjusted R^2, across all or many of the possible subset mod-

els. Indeed, as the number of covariates grows, searching through all possible subset models becomes prohibitive.

Variable selection remains an important and challenging issue in the rapidly growing realm of high-dimensional statistical modeling. Modern methods, increasingly based on the principle of penalized least squares (or negative loglikelihood) estimation applied to high dimensional regression problems, attempt to achieve this goal through an adaptive variable selection process that simultaneously permits estimation of regression effects.

Penalty functions that are nonsmooth (i.e. not differentiable) at the origin have received substantial attention. As noted in Fan and Li (2001), for example, there are many connections between thresholding rules and variable selection for linear models when the columns of the design matrix $\mathbf{X}$ are orthonormal. In such cases, the penalized least squares problem can be expressed in the form

$$\frac{1}{2}\sum_{i=1}^{N}(z_i-\theta_j)^2+\sum_{j=1}^{p}\tilde{p}_j(|\theta_j|;\lambda), \tag{2.1}$$

where the penalty functions $\tilde{p}_j(\cdot;\lambda)$ are not necessarily the same for all j, but often assumed so for simplicity; i.e., $\tilde{p}_j(\cdot;\lambda)=\tilde{p}(\cdot;\lambda)$ with λ as the associated penalty parameter. When the objective is of form (2.1), its minimization can be considered componentwise for each $j=1,\ldots,p$:

$$\frac{1}{2}(z_j-\theta_j)^2+\tilde{p}(|\theta_j|;\lambda).$$

For example, the L_1 penalty $\tilde{p}(|\theta|;\lambda)=\lambda|\theta|$ yields the soft thresholding rule

$$\hat{\theta}_j=sign(z_j)(|z_j|-\lambda)_+,\ \ j=1,\ldots,p, \tag{2.2}$$

proposed by Donoho and Johnstone (1994), where $(a)_+=\max(0,a)$. For the general (non-orthonormal) least squares setting, the LASSO estimator of Tibshirani (1996) is the penalized least squares estimate with the L_1 penalty.

Fan and Li (2001) argue that 'good' penalties should result in estimators with three properties: unbiasedness, sparsity, and continuity. Unbiasedness refers to the estimator being nearly unbiased when the true unknown parameter is large in magnitude. Sparsity implies that the estimator results from a thresholding rule, i.e., coefficients with small magnitude below a certain level are set to zero. Continuity in the observed z_j is desirable to avoid instability in the model prediction. With such specifications, a penalty function satisfying both the properties of continuity and sparsity must be nonsmooth (singular) at the origin (Fan and Li, 2001).

The bulk of the current literature tends to focus on specific combinations of smooth data fidelity (i.e., goodness-of-fit such as least squares criterion or negative loglikelihood functions) and nonsmooth penalty functions. One result of this combined specificity has been a proliferation in the number of computational algorithms designed to solve fairly narrow classes of optimization problems involving objective functions that are not everywhere continuously differentiable; see, for example, Table 2.1. Chapters 3 and 4 develop a unified optimization framework that appeals to the Majorization-Minimization (MM) algorithm (Lange et al., 2000; Lange, 2004) and accommodates the lack of everywhere differentiability. Specifically, these chapters focus on optimization of penalized objective functions of the form

$$\xi(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + p(\boldsymbol{\beta}; \boldsymbol{\lambda}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2, \;\; \lambda > 0, \; \varepsilon \geq 0, \tag{2.3}$$

where $\|\cdot\|$ denotes the usual Euclidean vector norm, and $g(\boldsymbol{\beta})$ and $p(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \sum_j \tilde{p}(\beta_j; \boldsymbol{\lambda}_j)$ are respectively data fidelity (e.g., negative loglikelihood) and nonsmooth penalty functions that satisfy regularity conditions described in Chapter 4. The class of problems represented by (2.3) contains all of the penalized regression problems commonly considered in the current literature and included in Table 2.1. It also covers numerous other problems by expanding the class of permissible fidelity and penalty functions in a substantial way.

Table 2.1: Different combinations of smooth data fidelity and nonsmooth penalty functions result in many optimization algorithms. Penalties (A)LAS, (A)EN, SCAD will be discussed in Chapter 4; these have been applied in linear models (LM), generalized linear models (GLM), Cox proportional hazards models (CPH), accelerated failure time models (AFT), and beyond.

	Penalty		
Data Fidelity	(A)LAS	(A)EN	SCAD
LM	Tibshirani (1996) Efron et al. (2004) Zou (2006)	Zou and Hastie (2005) Zou and Zhang (2009)	Fan and Li (2001) Hunter and Li (2005) Zou and Li (2008) Kim et al. (2008)
GLM	Park and Hastie (2007) Friedman et al. (2008)	Park and Hastie (2007) Friedman et al. (2008)	Fan and Li (2001) Hunter and Li (2005) Zou and Li (2008)
CPH	Tibshirani (1997) Park and Hastie (2007) Zhang and Lu (2007)	Park and Hastie (2007) Engler and Li (2009)	Fan and Li (2002) Hunter and Li (2005)
AFT	Huang et al. (2006) Datta et al. (2007) Cai et al. (2009)	Wang et al. (2008) Engler and Li (2009)	Johnson et al. (2008)

2.1.1 Penalized Finite Mixture Regression Models

It is often the case, especially with a large number of covariates, that the N observed samples are not adequately modeled using the same set of regression coefficients; that is, a set or subset of coefficients may be different for different subgroups of observations. Recently, Städler et al. (2010) considered L_1-penalized (linear) finite mixture regression (FMR) models for high dimensional data for a fixed number of components. The inclusion of the L_1 penalty induces the desirable property of sparsity in the coefficients. Their generalized EM algorithm is based explicitly on Block Coordinate Descent (BCD). The methods in Chapter 5 utilize the MM approach developed in Chapter 4, and can be considered as extensions of the Städler et al. (2010) framework to a broader class of regression models and penalty functions.

2.1.2 Hierarchical Modeling of the Penalty Parameter

Also recently in the literature, Park and Casella (2008) explored a Bayesian treatment of the L_1 constrained least squares regression problem. In such formulations, the penalty function is typically regarded as the negative logarithm of the coefficient prior distribution. For example, it is well-known that the LASSO estimate for the linear regression coefficients can be interpreted as a Bayesian posterior mode, or maximum a posteriori (MAP) estimate, when the coefficients have independent double exponential priors (Tibshirani, 1996). However, Park and Casella (2008) proposed a hierarchical model with conjugate normal priors for the regression coefficients and independent exponential priors on their variances to exploit the double exponential representation as a scale mixture of normals with an exponential mixing density. Generalizations with different mixing distributions have also been explored in Griffin and Brown (2005, 2007); Carvalho et al. (2010). The structure of the Park and Casella (2008) model allows for some uniquely Bayesian alternatives for selecting the Bayesian LASSO tuning parameter. Particularly of interest was their suggestion to place a diffuse hyperprior (class of gamma priors) on the squared tuning parameter, λ^2, and to use the posterior median as an estimate of λ.

Strawderman and Wells (2010) take a slightly different Bayesian perspective. They explore the connections between the hierarchical priors of Strawderman (1971) and Takada (1979), and their respective proper Bayes and MAP estimators in the multivariate normal means problem ($\mathbf{Z} \sim N_p(\boldsymbol{\theta}, \mathbf{I}_p)$). In particular, Strawderman and Wells (2010) consider maximizing the posterior distribution in both the mean vector $\boldsymbol{\theta}$ and hyperparameter λ under a specific choice of joint prior distribution $\pi(\boldsymbol{\theta}, \lambda|\alpha, \beta)$ where $\alpha, \beta > 0$ are hyperparameters.

The aforementioned works were the primary motivation behind the hierarchical model proposed in Chapter 6; for $p = 1$ and an appropriate choice of joint prior, we re-

cover the univariate Minimax Concave Penalty (MCP) thresholding estimator of Zhang (2010) as the MAP estimator.

2.2 MM algorithm

The MM algorithm is an iterative optimization method that, in essence, substitutes a difficult optimization problem with a simpler one. Indeed, the term "optimization transfer" was originally used to describe such algorithms in the the seminal work by Lange et al. (2000). When dealing with minimization, the substitution is with a majorizing function; hence MM stands for Majorize-Minimize. The majorizing function must be specially designed so that minimum of the majorizer coincides with the minimum of the desired objective function. In particular, let $\xi(\boldsymbol{\beta})$ denote a real-valued objective function to be minimized for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ in some convex subset $\mathcal{B}$ of $\mathbb{R}^p$. Let $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ denote a real-valued "surrogate" objective function, where $\boldsymbol{\alpha} \in \mathcal{B}$. Define the minimization map

$$M(\boldsymbol{\alpha}) = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\arg\min} \ \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha}). \tag{2.4}$$

Then, if $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ majorizes $\xi(\boldsymbol{\beta})$ for each $\boldsymbol{\alpha}$, i.e.,

$$\xi(\boldsymbol{\beta}) = \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta}) \text{ for each } \boldsymbol{\beta} \in \mathcal{B} \text{ and} \tag{2.5}$$

$$\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \geq \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta}) \text{ for } \boldsymbol{\beta} \neq \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha} \in \mathcal{B}, \tag{2.6}$$

a generic MM algorithm for minimizing $\xi(\boldsymbol{\beta})$ takes the form (e.g., Lange, 2004):

1. Initialize $\boldsymbol{\beta}^{(0)}$.
2. For $k \geq 0$, compute $\boldsymbol{\beta}^{(k+1)} = M(\boldsymbol{\beta}^{(k)})$, iterating until convergence.

Since $\boldsymbol{\beta}^{(k+1)}$ is the minimum of the surrogate function at $\boldsymbol{\beta}^{(k)}$, the MM procedure forces $\xi(\boldsymbol{\beta})$ downhill, i.e., $\xi(\boldsymbol{\beta}^{(k+1)}) \leq \xi(\boldsymbol{\beta}^{(k)})$. Provided that the objective function, its surrogate, and the mapping $M(\cdot)$ satisfy certain regularity conditions, one can establish convergence of this algorithm to a local or global solution. Lange (2004, Ch. 10) summarizes such a theory assuming that the objective functions $\xi(\boldsymbol{\beta})$ and $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are twice continuously differentiable. Drawing on results from Lange (1995), Lange et al. (2000) remark that global convergence for minimization requires the iteration map $M(\cdot)$ to be continuous and satisfy $\xi(M(\boldsymbol{\beta})) \leq \xi(\boldsymbol{\beta})$ with equality if and only if $\boldsymbol{\beta}$ is a fixed point of $M(\boldsymbol{\beta})$. Under the further assumption that the set of stationary points of $M(\boldsymbol{\beta})$ coincides with the set of stationary points of $\xi(\boldsymbol{\beta})$, it can be shown that any limit point of the sequence $\boldsymbol{\beta}^{(k+1)} = M(\boldsymbol{\beta}^{(k)})$ is a stationary point of $\xi(\boldsymbol{\beta})$. However, the notion of stationarity is not clearly defined within Lange et al. (2000), and seems to implicitly rely on differentiability. Later, in a different (specific) context, Lange et al. (2000) comment without much detail that optimization transfer "works" without differentiability when using an appropriately defined subdifferential. As most penalized objective functions are singular at the origin (e.g., (2.3)), the case of not-everywhere differentiability clearly deserves more attention.

2.2.1 MM and Penalized Likelihoods

While the MM algorithm has been considered previously for optimization of such non-smoothly penalized likelihood problems, a more complete and general theory of local convergence has been lacking. Hunter and Li (2005) and Zou and Li (2008), appearing a few times in Table 2.1, both make use of the MM algorithm and pay specific attention to solving broader classes of penalized estimation problems with a unified computational methodology. In particular, Hunter and Li (2005) developed a Local Quadratic Ap-

proximation (LQA) algorithm, which solves a perturbed (i.e., differentiable) version of desired optimization problem. Here, the concave, nondifferentiable penalty is replaced by differentiable approximation (local quadratic majorization) of the penalty. In contrast, Zou and Li (2008) solve an optimization problem involving a twice-differentiable data fidelity function and a concave, nondifferentiable penalty using the Local Linear Approximation (LLA) algorithm, so-named for the use of a local linear majorization of the penalty. In comparison, Hunter and Li (2005) establish stronger convergence results, facilitated by the differentiability of perturbed penalty. Zou and Li (2008) cite sensitivity of LQA to the choice of perturbation and its resulting impact on solution sparsity. However, as noted in Mazumder et al. (2009), Zou and Li (2008) never proved that LLA iteration sequence actually converges.

Earlier work also includes Daubechies et al. (2004), who consider penalized linear regression in a Hilbert space with the (convex) LASSO penalty. Of importance, they showed how a specific choice of majorization for the least squares objective function leads to an iterative algorithm that relies only on soft-thresholding, as well as prove convergence of the resulting iteration sequence (relying heavily on convexity).

The methods developed in Chapter 4 essentially hybridize the ideas of Zou and Li (2008) and Daubechies et al. (2004) into a numerically stable and versatile class of MM algorithms capable of dealing with a wide variety of penalized objective functions. Furthermore, the theory presented in Chapter 3 establishes convergence of the MM iteration sequence under weaker assumptions than existing work in this area. Related results for the EM algorithm have been established in Wu (1983), Tseng (2004) and Chrétien and Hero (2008), and are discussed further in Chapter 3.

2.2.2 Relation to EM algorithm

Perhaps the most well-known type of MM algorithm is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977), where in the maximization context, MM refers to Minorization-Maximization. The EM algorithm is a general approach for finding maximum likelihood estimates that relies on the concept of incomplete or missing data, where the incompleteness can be real or artificial. One typically posits a model for the unobserved complete data, which describes the problem of interest in the absence of missing data. Ideally, the model for the complete data is relatively simple to optimize. For example, finite mixture models are often fit with the EM algorithm, where group/component membership acts as the missing data, but the complete data is modeled as if the group/component membership was known. Indeed, this approach to finite mixture modeling is used in Chapter 5.

In the general EM paradigm, one typically wishes to maximize a loglikelihood $\ell(\boldsymbol{\phi})$ in the unknown parameters $\boldsymbol{\phi}$. Equivalently, one could minimize $g(\boldsymbol{\phi}) \equiv -\ell(\boldsymbol{\phi})$. Let $k(\mathbf{m}|\mathbf{o};\boldsymbol{\phi}) = f_c(\mathbf{m};\boldsymbol{\phi})/f(\mathbf{o};\boldsymbol{\phi})$ be the conditional density of the missing data $\mathbf{m}$ given the observed data $\mathbf{O} = \mathbf{o}$, where $f_c(\mathbf{m};\boldsymbol{\phi})$ represents the complete data likelihood. Then

$$\begin{aligned} g(\boldsymbol{\phi}) \equiv -\ell(\boldsymbol{\phi}) &= -\ell_C(\boldsymbol{\phi}) + \log k(\mathbf{m}|\mathbf{o};\boldsymbol{\phi}) \\ &= E_{\boldsymbol{\phi}^{(k)}}\{-\ell_C(\boldsymbol{\phi})|\mathbf{o}\} + E_{\boldsymbol{\phi}^{(k)}}\{\log k(\mathbf{M}|\mathbf{o};\boldsymbol{\phi})|\mathbf{o}\} \\ &= Q(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)}) + H(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)}), \end{aligned} \tag{2.7}$$

(e.g., McLachlan and Krishnan, 2008). Note that in the minimization context, the EM algorithm is a descent algorithm:

$$\begin{aligned} g(\boldsymbol{\phi}^{(k+1)}) - g(\boldsymbol{\phi}^{(k)}) &= Q(\boldsymbol{\phi}^{(k+1)},\boldsymbol{\phi}^{(k)}) - Q(\boldsymbol{\phi}^{(k)},\boldsymbol{\phi}^{(k)}) \\ &\quad + H(\boldsymbol{\phi}^{(k+1)},\boldsymbol{\phi}^{(k)}) - H(\boldsymbol{\phi}^{(k)},\boldsymbol{\phi}^{(k)}) \\ &\leq 0. \end{aligned}$$

This follows because (i) by definition $\boldsymbol{\phi}^{(k+1)}$ minimizes $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$ in $\boldsymbol{\phi}$ so $Q(\boldsymbol{\phi}^{(k+1)}, \boldsymbol{\phi}^{(k)}) \leq Q(\boldsymbol{\phi}^{(k)}, \boldsymbol{\phi}^{(k)})$, and (ii) $H(\boldsymbol{\phi}^{(k+1)}, \boldsymbol{\phi}^{(k)}) - H(\boldsymbol{\phi}^{(k)}, \boldsymbol{\phi}^{(k)}) \leq 0$ using Jensen's inequality (e.g., McLachlan and Krishnan, 2008). In fact, $H(\boldsymbol{\phi}, \boldsymbol{\theta}) \leq H(\boldsymbol{\theta}, \boldsymbol{\theta})$ for all $\boldsymbol{\phi}, \boldsymbol{\theta}$ in the parameter space $\boldsymbol{\Phi}$. Thus, minimizing $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) = g(\boldsymbol{\phi}) - H(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$ in $\boldsymbol{\phi}$ is equivalent to minimizing

$$g(\boldsymbol{\phi}) + D(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}),$$

in $\boldsymbol{\phi}$ where $D(\boldsymbol{\phi}, \boldsymbol{\theta}) = H(\boldsymbol{\theta}, \boldsymbol{\theta}) - H(\boldsymbol{\phi}, \boldsymbol{\theta})$ is nonnegative for all $\boldsymbol{\phi}, \boldsymbol{\theta} \in \boldsymbol{\Phi}$ and equals zero if $\boldsymbol{\phi} = \boldsymbol{\theta}$ (e.g., Tseng, 2004). Thus, $g(\boldsymbol{\phi})$ is majorized through the E-step by $g(\boldsymbol{\phi}) + D(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$; this is the first "M". The minimization of the majorizing function gives us the second "M".

2.3 The Clarke Subdifferential and its Properties

The traditional notion of a stationary point is defined using the gradient; that is, $\boldsymbol{\beta}^*$ is a stationary point of $\xi(\boldsymbol{\beta})$ if $\nabla\xi(\boldsymbol{\beta}) = \mathbf{0}$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. In problems where the gradient does not necessarily exist everywhere, the notion of a stationary point can be generalized using the theory of the Clarke subdifferential (Clarke, 1990). In particular, for a function $f : \mathbb{R}^p \to \mathbb{R}$ that is locally Lipschitz continuous and with $\partial f(\mathbf{x})$ denoting its corresponding Clarke subdifferential at a point $\mathbf{x}$, one defines $\mathbf{x}^*$ as a stationary point of f if $0^{p\times 1} \in \partial f(\mathbf{x}^*)$. The Clarke subdifferential, defined in C5 below, is analogous to the subdifferential of convex function theory (e.g., Hiriart-Urruty and Lemaréchal, 1996) and is a set. This generalized notion of stationarity is in fact a necessary condition for achieving stationarity in the sense of the the traditional definition and it does not require the existence of the gradient $\nabla\xi(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ to in order to assert the existence of a stationary point. However, if this gradient exists, then $\partial\xi(\boldsymbol{\beta}^*) = \{\nabla\xi(\boldsymbol{\beta}^*)\} = \{\mathbf{0}\}$.

For convenience, several key definitions and results from the theory of subdifferentials and nonsmooth optimization are now reviewed below; see, for example, Clarke (1990), Mäkelä and Neittaanmaäki (1992), and Melkonyan (2010).

C1. A function $f : \mathbb{R}^p \to \mathbb{R}$ is lower semicontinuous if the level sets $L(c) = \{b \in \mathbb{R}^p : f(b) \leq c\}$ are closed for each $c \in \mathbb{R}$. If f is lower semicontinuous, then $-f$ is upper semicontinuous. A function that is both lower semicontinuous and upper semicontinuous must be continuous. Suppose the sets defined above are also bounded; then, such functions that are lower (upper) semicontinuous are also lower (upper) compact (closed and bounded is equivalent to compact in $\mathbb{R}^p$).

C2. A function $f : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz continuous at x^* with constant L if there exists an $L \in [0, \infty)$ and $\rho > 0$ such that $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \{r \in \mathbb{R}^p : \|r - x^*\| \leq \rho\}$.

C3. If $f : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz continuous for $x \in \Omega \subset \mathbb{R}^p$, then $\nabla f(x)$ exists for almost all $x \in \Omega$.

C4. If $f : \mathbb{R}^p \to \mathbb{R}$ is convex for $x \in \Omega \subset \mathbb{R}^p$, it is locally Lipschitz continuous for $x \in \Omega \subset \mathbb{R}^p$.

C5. Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz continuous. The (Clarke) subdifferential of f at x is the set

$$\partial f(x) := \{v \in \mathbb{R}^p : f^o(x, d) \geq v^T d \text{ for all } d \in \mathbb{R}^p\},$$

where

$$f^o(x, d) = \limsup_{y \to x, t \to 0+} t^{-1} \big(f(y + td) - f(y)\big)$$

is the generalized directional derivative of f at x in the direction d. Each element $g \in \partial f(x)$ is referred to as a subgradient of f.

C6. Suppose $f : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz continuous at x. Then, f is said to be regular at x if the ordinary (Gâteaux) directional derivative

$$f'(x,d) = \lim_{t \to 0+} t^{-1} \left(f(x+td) - f(x) \right)$$

exists for all directions d and agrees with $f^o(x,d)$. Sufficient conditions for regularity include continuous differentiability or convexity.

C7. If $f_i : \mathbb{R}^p \to \mathbb{R},\ i = 1,2$ are locally Lipschitz continuous at x, then

$$\partial[f_1(x) + f_2(x)] \subset \partial f_1(x) \oplus \partial f_2(x),$$

where the set operation $A \oplus B$ denotes the set formed by adding every element of A to every element of B. Equality holds in the sense of set equivalence if at least one of $\partial f_i(x)$ is a singleton set (e.g., one of the functions is continuously differentiable); or, it holds if both f_1 and f_2 are regular.

C8. Suppose $f_1 : \mathbb{R}^p \to \mathbb{R}$, and $f_2 : \mathbb{R}^p \to \mathbb{R}^p$ are locally Lipschitz continuous at x. Suppose $f_1(s)$ is continuously differentiable at $s = f_2(x)$. Then, $\partial f_1(f_2(x)) = f_1'(f_2(x)) \times \partial f_2(x)$, where the set operation denotes multiplying each element of $\partial f_2(x)$ by the scalar $f_1'(f_2(x))$.

C9. If $f : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz continuous at x and also differentiable at x, then $\nabla f(x) \in \partial f(x)$.

C10. If $f : \mathbb{R}^p \to \mathbb{R}$ is continuously differentiable at x, it is locally Lipschitz continuous at x and $\partial f(x)$ reduces to the singleton set $\{\nabla f(x)\}$.

C11. If $f : \mathbb{R}^p \to \mathbb{R}$ is differentiable, regular, and locally Lipschitz at x, then $\partial f(x)$ reduces to the singleton set $\{\nabla f(x)\}$.

C12. If $f : \mathbb{R}^p \to \mathbb{R}$ is locally Lipschitz continuous at x and $f(x)$ is a local minimum, then $0^{p \times 1} \in \partial f(x)$ and $f^o(x,d) \geq 0$ for each $d \in \mathbb{R}^p$.

C13. Let $f : \mathbb{R}^p \to \mathbb{R}$. The Dini lower and upper directional derivatives are respectively defined as

$$\begin{aligned} f'_{D-}(x,d) &= \liminf_{t\to 0+} t^{-1}\left(f(x+td)-f(x)\right) \\ f'_{D+}(x,d) &= \limsup_{t\to 0+} t^{-1}\left(f(x+td)-f(x)\right). \end{aligned}$$

C14. If $f : \mathbb{R}^p \to \mathbb{R}$ has a Gâteaux directional derivative at x in the direction d then

$$f'_{D-}(x,d) = f'_{D+}(x,d) = f'(x,d).$$

C15. If $f : \mathbb{R}^p \to \mathbb{R}$ is convex, then $f'(x,d)$ exists for each d and

$$f'(x,d) = \inf_{t>0} t^{-1}\left(f(x+td)-f(x)\right).$$

C16. Let $f : \mathbb{R}^p \to \mathbb{R}$. f is Gâteaux differentiable at x if the limit

$$\lim_{t\to 0} t^{-1}\left(f(x+td)-f(x)\right)$$

exists for all d.

C17. If $f : \mathbb{R}^p \to \mathbb{R}$ is differentiable at x then $f'(x,d) = \left\langle \nabla f(x), d\right\rangle$.

CHAPTER 3

GENERAL MM LOCAL CONVERGENCE THEORY

As discussed in Chapter 2, a more complete and general local convergence theory is required for problems that are not everywhere differentiable, such as those of the form (2.3). Using convergence theory for algorithms derived from point-to-set maps developed by Zangwill (1969), Wu (1983) established the convergence of the EM algorithm assuming twice differentiability of the loglikelihood function. In particular, he showed that the limit points of the algorithm are stationary points of the loglikelihood, and under more stringent conditions, showed that the EM sequence is convergent. In what follows, the key convergence result of Zangwill (1969) is restated; this result, given in Theorem 3.1.1 and adapted from Wu (1983), is stated in a form convenient for use with the MM algorithm and provides for a very general (and comparatively weak) form of convergence. Drawing on stronger convergence results due to Meyer (1976), a more useful convergence theory for MM algorithms designed to minimize nondifferentiable objective functions is established; this result is stated in Theorem 3.1.3. Finally, a set of sufficient regularity conditions is provided that ensures the validity of the conditions of both theorems in a wide class of statistical estimation problems.

3.1 Convergence of MM Algorithms in Nonsmooth Problems

Let $\xi(\boldsymbol{\beta})$ be the real-valued function to be minimized, where $\boldsymbol{\beta} \in \mathcal{B}$ and $\mathcal{B}$ is some convex subset of $\mathbb{R}^p$. Let $M : \mathcal{B} \to \mathcal{B}$ be the minimization map (2.4), where $\xi^{SUR}(\cdot, \cdot)$ is any function that majorizes $\xi(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathcal{B}$. In general, $M(\cdot)$ is a point-to-set map, and therefore a set. $\bar{\boldsymbol{\beta}}$ is a *generalized fixed point* of $M(\cdot)$ if $\bar{\boldsymbol{\beta}} \in M(\bar{\boldsymbol{\beta}})$; $\bar{\boldsymbol{\beta}}$ is a *fixed point* of $M(\cdot)$ if $M(\bar{\boldsymbol{\beta}}) = \{\bar{\boldsymbol{\beta}}\}$ (i.e., a singleton). The main result of Zangwill (1969, Theorem A), also utilized in Wu (1983), is stated below.

Theorem 3.1.1. *Suppose $\xi(\boldsymbol{\beta})$ is a continuous, real-valued function of $\boldsymbol{\beta} \in \mathcal{B}$ that is uniformly bounded below. Let $\mathcal{S} \subset \mathcal{B}$ denote the (nonempty) set of stationary points (in the sense of Clarke (1990) of $\xi(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathcal{B}$ and assume the sequence $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ is generated as follows:*

- *$\boldsymbol{\beta}^{(0)} \in \mathcal{B}$, where $\boldsymbol{\beta}^{(0)}$ and $\xi(\boldsymbol{\beta}^{(0)})$ are bounded;*
- *$\boldsymbol{\beta}^{(k+1)} \in M(\boldsymbol{\beta}^{(k)})$, where $M(\cdot)$ is the point-to-set map* (2.4)*.*

Suppose that

Z1. Each $\boldsymbol{\beta}^{(k)} \in \mathcal{B}_0$, where the compact set $\mathcal{B}_0 \subset \mathcal{B}$;

Z2. $M(\cdot)$ is closed and non-empty for $\boldsymbol{\beta} \in \mathcal{S}^c \cap \mathcal{B}_0$.

Z3. We have:

(i) $\xi(\boldsymbol{\beta}) \leq \xi(\boldsymbol{\alpha})$ for each $\boldsymbol{\alpha} \in \mathcal{S}$ and any $\boldsymbol{\beta} \in M(\boldsymbol{\alpha})$;

(ii) $\xi(\boldsymbol{\beta}) < \xi(\boldsymbol{\alpha})$ for each $\boldsymbol{\alpha} \notin \mathcal{S}$ and any $\boldsymbol{\beta} \in M(\boldsymbol{\alpha})$.

Then, the following conclusions hold:

M1. The sequence $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ has at least one limit point in $\mathcal{S}$, and the set of all limit points, say $\mathcal{S}_0$, satisfies $\mathcal{S}_0 \subseteq \mathcal{S}$;

M2. Each limit point $\bar{\boldsymbol{\beta}} \in \mathcal{S}_0$ satisfies $\lim_{k\to\infty} \xi(\boldsymbol{\beta}^{(k)}) = \xi(\bar{\boldsymbol{\beta}})$.

M3. Each limit point $\bar{\boldsymbol{\beta}} \in \mathcal{S}_0$ is a generalized fixed point of $M(\cdot)$.

Remark 3.1.2. *Assumptions [Z1]-[Z3] are imposed in Wu (1983). Assumption [Z1] implies $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ is a bounded sequence, ensuring the existence of at least one limit point. Comments on [Z2] will be made below, as it is possible to impose reasonable*

sufficient conditions that ensure this condition. Assumption [Z3] enforces the descent property at each update, as would be expected in any EM, GEM or MM algorithm. An equivalent formulation of this condition follows (e.g. Meyer, 1976, p. 114):

Z3′. For each $\boldsymbol{\alpha} \in \mathcal{B}_0$ and $\boldsymbol{\beta} \in M(\boldsymbol{\alpha})$:

(i) $\xi(\boldsymbol{\beta}) < \xi(\boldsymbol{\alpha})$ if $\boldsymbol{\alpha} \notin M(\boldsymbol{\alpha})$ (i.e., a strict decrease occurs at points $\boldsymbol{\alpha}$ that are not generalized fixed points);

(ii) $\xi(\boldsymbol{\beta}) \leq \xi(\boldsymbol{\alpha})$ if $\boldsymbol{\alpha} \in M(\boldsymbol{\alpha})$ (i.e., if $\boldsymbol{\alpha}$ is a generalized fixed point, it is possible to observe no change in the objective function).

3.1.1 Convergence of the Iteration Sequence

The above theorem essentially guarantees convergence of subsequences, but not global convergence of the iteration sequence itself. Subsequential convergence permits, for example, oscillatory behavior in the limit sequence. Meyer (1976, 1977) offers several refinements of Theorem 3.1.1, strengthening the statements of convergence. His results, adapted for the MM algorithm, follow below; in particular, see Theorems 3.1, 3.5, 3.6 and Corollary 3.2 of Meyer (1976).

Theorem 3.1.3. *Let the conditions of Theorem 3.1.1 hold. Consider the following two additional conditions:*

Z4. For each $\boldsymbol{\alpha} \in \mathcal{B}_0$ and any $\boldsymbol{\beta} \in M(\boldsymbol{\alpha})$, we have $\xi(\boldsymbol{\beta}) < \xi(\boldsymbol{\alpha})$ whenever $M(\boldsymbol{\alpha}) \neq \{\boldsymbol{\alpha}\}$ (i.e., a strict decrease in the objective function occurs at any point $\boldsymbol{\alpha}$ that is not a fixed point);

Z5. there exists an isolated limit point $\bar{\boldsymbol{\beta}}^*$ such that $M(\bar{\boldsymbol{\beta}}^*) = \{\bar{\boldsymbol{\beta}}^*\}$ (i.e., a true fixed point).

Suppose [Z1]-[Z4] hold. Then, in addition to results [M1]-[M3] of Theorem 3.1.1, the following conclusions hold:

M4. *Each limit point $\bar{\boldsymbol{\beta}} \in \mathcal{S}_0$ satisfies $M(\bar{\boldsymbol{\beta}}) = \{\bar{\boldsymbol{\beta}}\}$, and is thus a fixed point of $M(\cdot)$;*

M5. $\lim_{k\to\infty} \|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k+1)}\| = 0$, *in which case one either has (i) the set of limit points $\mathcal{S}_0$ consists of a single point to which $\boldsymbol{\beta}^{(k)}$ converges; or, (ii) the set of limit points $\mathcal{S}_0$ forms a continuum, and $\boldsymbol{\beta}^{(k)}$ fails to converge;*

M6. *If the number of fixed points having any given value of $\xi(\cdot)$ is finite, then $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ converges to one of these fixed points;*

M7. *If the sequence $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ has an isolated fixed point $\bar{\boldsymbol{\beta}}$, then $\boldsymbol{\beta}^{(k)} \to \bar{\boldsymbol{\beta}}$. If $\bar{\boldsymbol{\beta}}$ is also an isolated local minimum of $\xi(\cdot)$ on $\mathcal{B}_0$, then there exists an open neighborhood $\mathcal{B}_\epsilon \subseteq \mathcal{B}_0$ of $\bar{\boldsymbol{\beta}}$ such that $\boldsymbol{\beta}^{(k)} \to \bar{\boldsymbol{\beta}}$ if $\boldsymbol{\beta}^{(0)} \in \mathcal{B}_\epsilon$.*

Suppose instead that [Z1-Z3] and [Z5] hold. Then, in addition to results [M1]-[M3] of Theorem 3.1.1, the following conclusion can be drawn:

M8. *If the sequence $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ has an isolated generalized fixed point $\bar{\boldsymbol{\beta}}$ that satisfies $M(\bar{\boldsymbol{\beta}}) = \{\bar{\boldsymbol{\beta}}\}$, then $\boldsymbol{\beta}^{(k)} \to \bar{\boldsymbol{\beta}}$. If $\bar{\boldsymbol{\beta}}$ is also an isolated local minimum of $\xi(\cdot)$ on $\mathcal{B}_0$, then there exists an open neighborhood $\mathcal{B}_\epsilon \subseteq \mathcal{B}_0$ of $\bar{\boldsymbol{\beta}}$ such that $\boldsymbol{\beta}^{(k)} \to \bar{\boldsymbol{\beta}}$ if $\boldsymbol{\beta}^{(0)} \in \mathcal{B}_\epsilon$.*

Remark 3.1.4. *Assumption [Z4] strengthens [Z3] by imposing the condition that the iteration scheme is* strictly *monotonic; as such, all generalized fixed points of $M(\cdot)$ are also fixed points, a situation that permits stronger statements of convergence results. Assumption [Z5] imposes the somewhat weaker assumption that there exists at least one isolated fixed point of the iteration sequence; similarly to [M7], [M8] implies that the iteration converges to this point.*

Conclusions [M1]-[M7] essentially mirror those in Vaida (2005, Theorems 1-3), who obtains strong convergence results for EM and MM algorithms under global differentiability assumptions on the objective and majorization functions and the additional condition that $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ has a unique global minimizer in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha} \in \mathcal{S}$, where $\mathcal{S}$ is a finite set of isolated stationary points. This uniqueness condition, encapsulated in [Z4], provides a verifiable condition for convergence of the MM algorithm that is often satisfied in statistical applications.

3.1.2 Sufficient Regularity Conditions for Local Convergence

Sufficient conditions that ensure [Z1]-[Z4], but weaker than those imposed in Vaida (2005), are now provided. In particular, suppose the objective function, its surrogate and the mapping $M(\cdot)$ satisfy the following regularity conditions:

R1. $\xi(\boldsymbol{\beta})$ is locally Lipschitz continuous and coercive for $\boldsymbol{\beta} \in \mathcal{B}$; that is, $L(\xi(\mathbf{z})) = \{\mathbf{b} \in \mathcal{B} : \xi(\mathbf{b}) \leq \xi(\mathbf{z})\}$ is compact for each $\mathbf{z} \in \mathcal{B}$. Consequently, $\xi(\boldsymbol{\beta})$ somewhere interior to $\mathcal{B}$; assume the elements of $\mathcal{S}$, the set of stationary points for $\xi(\boldsymbol{\beta})$, are isolated.

R2. $\xi(\boldsymbol{\beta}) = \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta})$ for each $\boldsymbol{\beta} \in \mathcal{B}$.

R3. $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha}) > \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta})$ for $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$, $\boldsymbol{\beta}, \boldsymbol{\alpha} \in \mathcal{B}$.

R4. $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is continuous for $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{B} \times \mathcal{B}$ and locally Lipschitz continuous in $\boldsymbol{\beta}$ for $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$.

R5. $M(\boldsymbol{\beta})$ exists and is a singleton set for each $\boldsymbol{\beta} \in \mathcal{B}$.

The above conditions do not imply that the objective function $\xi(\boldsymbol{\beta})$ is differentiable everywhere. Condition R1 does imply that $\xi(\boldsymbol{\beta})$ is bounded for $\boldsymbol{\beta}$ interior to $\mathcal{B}$ and that

$\nabla\xi(\boldsymbol{\beta})$ exists for almost all $\boldsymbol{\beta}$. Condition R1 further implies that the set of stationary points $\mathcal{S}$ is finite, as an infinite number of stationary points on a compact set would admit a convergent sequence whose limit would not be isolated. Conditions R2 and R3 imply that $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ strictly majorizes $\xi(\boldsymbol{\beta})$ and, in addition,

$$\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \xi(\boldsymbol{\beta}) + \psi(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{3.1}$$

where $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) := \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha}) - \xi(\boldsymbol{\beta})$ satisfies $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 0$ for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ and $\psi(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$. Conditions R4 and R5 imply that the map $M(\boldsymbol{\beta})$ is continuous, hence bounded on compact sets (Polak, 1987, Prop. 3.2). Conditions R1, R4, and R5 further imply that (3.1) is bounded below for $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{B} \times \mathcal{B}$ and that $\psi(\bar{\boldsymbol{\beta}}, \boldsymbol{\alpha})$ is uniquely minimized at $\boldsymbol{\alpha} = \bar{\boldsymbol{\beta}}$ for any fixed point $\bar{\boldsymbol{\beta}}$.

Suppose conditions R1-R5 hold. As commented earlier, conditions R4 and R5 imply that $M(\boldsymbol{\beta})$ is a continuous point-to-point map; hence, $M(\cdot)$ is closed (e.g. Luenberger and Ye, 2008, pp. 203-204), establishing [Z2]. Propositions 3.1.5 and 3.1.6, given below and proved under conditions R1-R5, now establish [Z1], [Z3] and [Z4].

Proposition 3.1.5. *Suppose $\boldsymbol{\beta}^{(k)}$ is bounded for a given $k \geq 0$. Then, $\boldsymbol{\beta}^{(k+1)} = M(\boldsymbol{\beta}^{(k)})$ exists, is bounded and is unique. In addition, for $k \geq 0$,*

$$\xi^{SUR}(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) \leq \xi^{SUR}(\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)}) < \infty \tag{3.2}$$

and

$$\xi(\boldsymbol{\beta}^{(k+1)}) - \xi(\boldsymbol{\beta}^{(k)}) \leq -\psi(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) \leq 0, \tag{3.3}$$

where the second inequality is strict unless $\boldsymbol{\beta}^{(k+1)} = M(\boldsymbol{\beta}^{(k)}) = \boldsymbol{\beta}^{(k)}$.

Proposition 3.1.6. *Let $\boldsymbol{\beta}^{(0)}$ be bounded. Define $\xi^{(k)} = \xi(\boldsymbol{\beta}^{(k)})$ for $k \geq 0$. Then, $\{\xi^{(k)}, k \geq 0\}$ is a bounded, monotone decreasing sequence. Moreover, the sequence $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ is bounded and contained in the compact set $L(\xi^{(0)})$.*

Proof of Proposition 3.1.5: Let $\boldsymbol{\alpha}$ be bounded but otherwise arbitrary. The continuity of $M(\cdot)$, along with condition R5, implies that $M(\boldsymbol{\alpha})$ exists, is bounded, and is unique. Using (2.4) and condition R2, we have that $\xi^{SUR}(M(\boldsymbol{\alpha}), \boldsymbol{\alpha}) \leq \xi^{SUR}(\boldsymbol{\alpha}, \boldsymbol{\alpha}) = \xi(\boldsymbol{\alpha}) < \infty$. Hence, (3.2) holds upon setting $\boldsymbol{\alpha} = \boldsymbol{\beta}^{(k)}$.

To establish (3.3), note that (3.1), (3.2) and the definition of $\boldsymbol{\beta}^{(k+1)}$ imply

$$\xi^{SUR}(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) = \xi(\boldsymbol{\beta}^{(k+1)}) + \psi(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) < \infty.$$

Using (3.2) and the fact that $\xi^{SUR}(\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)}) = \xi(\boldsymbol{\beta}^{(k)}) + \psi(\boldsymbol{\beta}^{(k)}, \boldsymbol{\beta}^{(k)}) = \xi(\boldsymbol{\beta}^{(k)})$, we further observe

$$\xi(\boldsymbol{\beta}^{(k+1)}) + \psi(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\beta}^{(k)}) \leq \xi(\boldsymbol{\beta}^{(k)}).$$

from which (3.3) is immediate. Under R3 and R4, this inequality is necessarily strict unless $\boldsymbol{\beta}^{(k+1)} = M(\boldsymbol{\beta}^{(k)}) = \boldsymbol{\beta}^{(k)}$, proving the result. □

Proof of Proposition 3.1.6: Since $\boldsymbol{\beta}^{(0)}$ is bounded, condition R1 implies $\xi^{(0)}$ is bounded, $\boldsymbol{\beta}^{(0)} \in L(\xi^{(0)})$, and $L(\xi^{(0)})$ is compact. From Proposition 3.1.5 and condition R5, we further observe that $\boldsymbol{\beta}^{(1)} = M(\boldsymbol{\beta}^{(0)})$ is bounded and satisfies $\boldsymbol{\beta}^{(1)} \in L(\xi^{(0)})$. Using condition R1 once more, $\xi^{(1)} = \xi(\boldsymbol{\beta}^{(1)})$ is bounded and, by (3.3), satisfies $\xi^{(1)} \leq \xi^{(0)}$; thus, $L(\xi^{(1)}) \subset L(\xi^{(0)})$.

We now use induction. Let $\boldsymbol{\beta}^{(k)}$ be bounded for some $k \geq 1$ and satisfy $\xi^{(k)} \leq \xi^{(0)}$; then, $\xi^{(k)}$ is necessarily bounded and $\boldsymbol{\beta}^{(k)} \in L(\xi^{(k)}) \subset L(\xi^{(0)})$. It again follows from Proposition 3.1.5 and condition R5 that $\boldsymbol{\beta}^{(k+1)} = M(\boldsymbol{\beta}^{(k)})$ is bounded and satisfies $\boldsymbol{\beta}^{(k+1)} \in L(\xi^{(k)})$. Hence, $\xi^{(k+1)}$ is bounded and satisfies $\xi^{(k+1)} \leq \xi^{(k)} \leq \xi^{(0)}$. Consequently, $L(\xi^{(k+1)}) \subset L(\xi^{(k)}) \subset L(\xi^{(0)})$ and $\boldsymbol{\beta}^{(k+1)} \in L(\xi^{(0)})$; it now follows that $\xi^{(k+1)} \leq \xi^{(k)}$, $L(\xi^{(k+1)}) \subset L(\xi^{(k)}) \subset L(\xi^{(0)})$, and $\boldsymbol{\beta}^{(k)} \in L(\xi^{(0)})$ for $k \geq 0$. Since $\xi(\cdot)$ is bounded below, $\{\xi^{(k)}, k \geq 0\}$ forms a bounded, monotone decreasing sequence and $\{\boldsymbol{\beta}^{(k)}, k \geq 0\}$ forms a bounded sequence contained within the compact set $L(\xi^{(0)})$. □

CHAPTER 4

MINIMIZATION BY ITERATIVE SOFT THRESHOLDING

Based on the theory summarized in Chapter 3, a new and general class of algorithms is proposed for minimizing penalized objective functions of the form (2.3):

$$\xi(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + p(\boldsymbol{\beta}; \boldsymbol{\lambda}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2, \;\; \lambda > 0, \; \varepsilon \geq 0$$

where $\|\cdot\|$ denotes the usual Euclidean vector norm, and $g(\boldsymbol{\beta})$ and $p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ are respectively data fidelity (e.g., negative loglikelihood) and penalty functions that satisfy regularity conditions to be delineated below. As will be shown later, the class of problems represented by (2.3) contains all of the penalized regression problems commonly considered in the current literature. It also covers numerous other problems by expanding the class of permissible fidelity and penalty functions in a substantial way.

We begin this chapter with a theorem providing sufficient conditions for the application of the general MM local convergence theory summarized in Chapter 3, and introduce the MM-based Minimization by Iterative Soft Thresholding (MIST) algorithm for minimizing (2.3). Simulation results and an application to survival data analysis follow; proofs are relegated to the end of this chapter.

4.1 MM Penalized Regression Formulation

Assume throughout that $g(\boldsymbol{\beta})$ is convex and coercive for $\boldsymbol{\beta} \in \mathcal{B}$; that is, the level set $L(g(\mathbf{z})) = \{\mathbf{b} \in \mathcal{B} : g(\mathbf{b}) \leq g(\mathbf{z})\}$ is compact for each $\mathbf{z} \in \mathcal{B}$, where $\mathcal{B}$ is an open convex subset of $\mathbb{R}^p$. Further assume that

$$p(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \sum_{j=1}^{p} \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j), \tag{4.1}$$

where the vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_p^T)^T$ and $\boldsymbol{\lambda}_j$ denotes the block of $\boldsymbol{\lambda}$ associated with β_j. It is assumed that each $\boldsymbol{\lambda}_j$ has dimension greater than or equal to one, that all blocks have the same dimension, and that the $\boldsymbol{\lambda}_{j1} = \lambda$ for each $j \geq 1$. Evidently, the case where $\dim(\boldsymbol{\lambda}_j) = 1$ for $j \geq 1$ simply corresponds to the setting in which each coefficient is penalized in exactly the same way; permitting the dimension of $\boldsymbol{\lambda}_j$ to exceed one allows the penalty to depend on additional parameters (e.g., weights, such as in the case of the adaptive lasso considered in Zou (2006)). We are interested in problems with penalization; therefore, λ is assumed bounded and strictly positive. Several specific examples will be discussed below. For any bounded $\boldsymbol{\theta}$ with $\lambda > 0$ as the first element, and the remainder of $\boldsymbol{\theta}$ collecting any additional parameters used to define the penalty, the scalar function $\tilde{p}(r; \boldsymbol{\theta})$ is assumed to satisfy the following condition:

(P1) $\tilde{p}(r; \boldsymbol{\theta}) > 0$ for $r > 0$; $\tilde{p}(0; \boldsymbol{\theta}) = 0$; $\tilde{p}(r; \boldsymbol{\theta})$ is a continuously differentiable concave function with $\tilde{p}'(r; \boldsymbol{\theta}) \geq 0$ for $r > 0$, and, $\tilde{p}'(0+; \boldsymbol{\theta}) \in [M_{\boldsymbol{\theta}}^{-1}, M_{\boldsymbol{\theta}}]$ for some finite $M_{\boldsymbol{\theta}} > 0$.

Notably, condition (P1) implies $\tilde{p}'(r; \boldsymbol{\theta}) > 0$ for $r \in (0, K_{\boldsymbol{\theta}})$, where $K_{\boldsymbol{\theta}} > 0$ may be finite or infinite. The setting in which (4.1) is identically zero for $r > 0$ is thus ruled out by the positivity of the right derivative at the origin imposed in (P1). This is not viewed as problematic, in that the specific interest lies in estimation subject to penalty singular at the origin. Moreover, were (4.1) absent, the convexity of $g(\boldsymbol{\beta})$ and strict convexity and continuous differentiability of the ridge-type penalty term $\lambda\varepsilon\|\boldsymbol{\beta}\|^2$ for $\varepsilon > 0$ implies (2.3) can be minimized directly using any suitable convex optimization algorithm.

Theorem 4.1.1 establishes local convergence of the indicated class of MM algorithms for minimizing objective functions of form (2.3). A proof is provided in Section 4.5, where it is shown that conditions imposed in the statement of the theorem are sufficient conditions for the application of the general theory summarized in Chapter 3.

Theorem 4.1.1. *Let $g(\boldsymbol{\beta})$ be convex and coercive and assume $p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ satisfies both* (4.1) *and condition (P1). Let $h(\boldsymbol{\beta}, \boldsymbol{\alpha}) \geq 0$ be a real-valued, continuous function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ that is continuously differentiable in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha}$ and satisfies $h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$ when $\boldsymbol{\beta} = \boldsymbol{\alpha}$. Let*

$$q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) = \sum_{j=1}^{p} \tilde{q}(|\beta_j|, |\alpha_j|; \boldsymbol{\lambda}_j), \tag{4.2}$$

where $\tilde{q}(r, s; \boldsymbol{\theta}) = \tilde{p}(s; \boldsymbol{\theta}) + \tilde{p}'(s; \boldsymbol{\theta})(r - s)$ for $r, s \geq 0$, and define

$$\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda}).$$

Assume the elements of $\mathcal{S}$, the set of stationary points for $\xi(\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathcal{B}$, are isolated. Then:

(i) $\xi(\boldsymbol{\beta})$ *in* (2.3) *is locally Lipschitz continuous and coercive;*

(ii) $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ *is either identically zero or non-negative for all $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$;*

(iii) $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \equiv \xi(\boldsymbol{\beta}) + \psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *majorizes $\xi(\boldsymbol{\beta})$ and the MM algorithm derived from $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ converges to a stationary point of $\xi(\boldsymbol{\beta})$ if $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is uniquely minimized in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha}$ and at least one of $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ or $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is strictly positive for each $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$.*

Condition (iii) of Theorem 4.1.1 establishes convergence under the assumption that $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ strictly majorizes $\xi(\boldsymbol{\beta})$ and has a unique minimizer in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha}$. Such a uniqueness condition is shown by Vaida (2005) to ensure convergence of the EM and MM algorithms to a stationary point under more restrictive differentiability conditions. Importantly, the assumption of globally strict majorization is only a sufficient condition for convergence; this condition is only important insofar as it guarantees a strict decrease in the objective function at every iteration. As can be seen from the proof in Section 4.5, it is possible to relax this condition to locally strict majorization, in which $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$

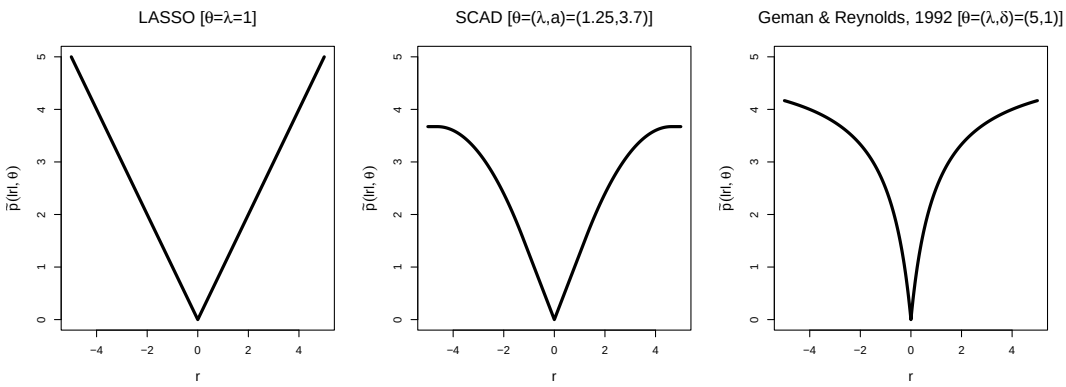

Figure 4.1: Three examples of penalties satisfying (P1).

majorizes $\xi(\boldsymbol{\beta})$, with strict majorization being necessary only in an open neighborhood containing $M(\boldsymbol{\alpha})$.

The use of the MM algorithm and selection of (4.2) are motivated by the results Zou and Li (2008); refer to Remark 4.2.1 below for further comments in this direction. The assumptions on $g(\boldsymbol{\beta})$ clearly cover the case of the linear and canonically parameterized generalized linear models upon setting $g(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$, where $\ell(\boldsymbol{\beta})$ denotes the corresponding loglikelihood function. Estimation under the semiparametric Cox regression model (Cox, 1972) and accelerated failure time models are also covered upon setting $g(\boldsymbol{\beta})$ to be either the negative logarithm of the partial likelihood function (e.g., Andersen et al., 1993, Thm VII.2.1) or the Gehan objective function (e.g., Fygenson and Ritov, 1994; Johnson and Strawderman, 2009).

The assumption (P1) on the penalty function covers a wide variety of popular and interesting examples; see Figure 4.1 for illustration. For example, the LASSO (LAS; e.g., Tibshirani, 1996), adaptive LASSO (ALAS; e.g., Zou, 2006), elastic net (EN; e.g., Zou and Hastie, 2005), and adaptive elastic net (AEN; e.g., Zou and Zhang, 2009) penalties are all recovered as special cases upon considering the combination of (4.1) and the ridge-type penalty $\lambda\varepsilon\|\boldsymbol{\beta}\|^2$. Specifically, with $\boldsymbol{\lambda}_j = (\lambda, \omega_j)^T$ for $\omega_j \geq 0$, taking

$\tilde{p}(r;\boldsymbol{\lambda}_j) = \lambda\omega_j r$ in (4.1) gives LAS ($\omega_j = 1, \varepsilon = 0$), ALAS ($\omega_j > 0, \varepsilon = 0$), EN ($\omega_j = 1, \varepsilon > 0$) and the AEN ($\omega_j > 0, \varepsilon > 0$) penalties. It is easy to see that selecting $\tilde{p}(r;\boldsymbol{\lambda}_j) = \lambda\omega_j r$ also implies the equality of (4.1) and (4.2), a result relevant in both (ii) and (iii) of Theorem 4.1.1 above.

The proposed penalty specification also covers the smoothly clipped absolute deviation (SCAD; e.g., Fan and Li, 2001) penalty upon setting $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_S(r;\lambda,a)$ for each $j \geq 1$, where $\tilde{p}_S(r;\lambda,a)$ is defined as the definite integral of

$$\tilde{p}'_S(u;\lambda,a) = \lambda[I(u \leq \lambda) + \frac{(a\lambda - u)_+}{(a-1)\lambda} I(u > \lambda)] \tag{4.3}$$

on the interval $0 \leq u \leq r$ and some fixed value of $a > 2$ (e.g., $a = 3.7$). The resulting penalty function is continuously differentiable and concave on $r \in [0,\infty)$. The concavity of $\tilde{p}_S(\cdot;\lambda,a)$ on $[0,\infty)$, combined with $\tilde{p}_S(0;\lambda,a) = 0$ and the fact that $\tilde{p}'_S(0+;\lambda,a)$ is finite, implies

$$\tilde{p}_S(r;\lambda,a) \leq \tilde{p}_S(s;\lambda,a) + \tilde{p}'_S(s;\lambda,a)(r-s) \tag{4.4}$$

for each $r, s \geq 0$, the boundary cases for $r = 0$ and/or $s = 0$ following from Hiriart-Urruty and Lemaréchal (1996, Remark 4.1.2, p. 21). In other words, $\tilde{p}_S(r;\lambda,a)$ can be majorized by a linear function of r.

The LASSO penalty, its variants, and SCAD have received the greatest attention in the literature. More recently, Zhang (2010) introduced the minimax concave penalty (MCP), which similarly to SCAD is defined in terms of its derivative. Specifically, one takes $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_M(r;\lambda,a)$ for each $j \geq 1$ in (4.1), where $\tilde{p}_M(r;\lambda,a)$ is defined as the definite integral of

$$\tilde{p}'_M(u;\lambda,a) = \left(\lambda - \frac{u}{a}\right)_+ \tag{4.5}$$

on the interval $0 \leq u \leq r$ and some fixed value of $a > 1$ (e.g., $a = 3.7$ as in Fan et al., 2009b). Further examples of differentiable concave penalties satisfying condition (P1)

include $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_G(r;\lambda,\delta)$ for

$$\tilde{p}_G(r;\lambda,\delta) = \lambda\frac{\delta r}{1+\delta r},\ \delta > 0 \tag{4.6}$$

(e.g., Geman and Reynolds, 1992; Nikolova, 2000); and $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_Y(r;\lambda,\delta)$ for

$$\tilde{p}_Y(r;\lambda,\delta) = \lambda\log(\delta r+1),\ \delta > 0; \tag{4.7}$$

(e.g., Antoniadis et al., 2009). These penalties represent just a small sample of the set of possible penalties satisfying (P1) that one might reasonably consider.

Remark 4.1.2. *The SCAD and MCP penalties lead to surrogate majorizers that fail to satisfy the globally strict majorization condition in (iii) of Theorem 4.1.1 unless $h(\boldsymbol{\beta},\boldsymbol{\alpha})$ is strictly positive whenever $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$; see, for example, Theorem 4.2.4 below.*

4.2 MIST: Minimization by Iterated Soft Thresholding

In general, the statistical literature on penalized estimation has proposed optimization algorithms tailored for specific combinations of fidelity and penalty functions. The class of MM algorithms suggested by Theorem 4.1.1 provides a very general and useful framework for proposing new algorithms, the key to which is a methodology for solving the minimization problem (2.4), a step repeated with each iteration of the MM algorithm. In this regard, it is helpful to note that the problem of minimizing $\xi^{SUR}(\boldsymbol{\beta},\boldsymbol{\alpha})$ for a given $\boldsymbol{\alpha}$ is equivalent to minimizing

$$g(\boldsymbol{\beta}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2 + h(\boldsymbol{\beta},\boldsymbol{\alpha}) + \sum_{j=1}^{p}\tilde{p}'(|\alpha_j|;\boldsymbol{\lambda}_j)|\beta_j| \tag{4.8}$$

in $\boldsymbol{\beta}$. In particular, if $g(\boldsymbol{\beta}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2 + h(\boldsymbol{\beta},\boldsymbol{\alpha})$ is strictly convex for each bounded $\boldsymbol{\alpha}$, which clearly occurs if both $g(\boldsymbol{\beta})$ and $h(\boldsymbol{\beta},\boldsymbol{\alpha})$ are convex in $\boldsymbol{\beta}$ and at least one is strictly convex, then (4.8) is also strictly convex and the corresponding minimization problem has a unique solution.

Remark 4.2.1. *For* $\varepsilon = h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$ *and* $g(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$ *for* $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta})$ *with* $\ell_i(\boldsymbol{\beta})$ *a twice continuously differentiable loglikelihood function, the MM algorithm induced by the surrogate function* (4.8) *corresponds (up to sign) to the minorizer employed in the LLA algorithm of Zou and Li (2008), an improvement upon the so-called LQA algorithm considered in Hunter and Li (2005). (Zou and Li, 2008, Proposition 1) assert convergence of their LLA algorithm under imprecisely stated assumptions and are additionally unclear as to the nature of convergence result actually established. For example, while (Zou and Li, 2008, Theorem 1) demonstrate that the LLA algorithm does indeed have an ascent property, their result appears to be insufficient to ensure that the proper analog of condition Z3(ii) of Theorem 3.1.1 holds in the case of the SCAD penalty. In particular, convergence of the LLA solution sequence is never actually established. In contrast, Theorem 4.1.1 shows that strict majorization, under a few precisely stated conditions, is sufficient to ensure local convergence of the resulting MM algorithm to a stationary point of* (2.3). *In Section 4.2.1, it is further demonstrated how a particular choice of* $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *yields a strict majorizer that permits both closed form minimization and componentwise updating at each step of the MM algorithm, even in the case of penalties that fail to be strictly concave.*

Numerous methods exist for minimizing a differentiable convex objective function (e.g., Boyd and Vandenberghe, 2004). However, because (4.8) is not differentiable, such methods do not apply in the current setting. Specialized methods exist for nonsmooth problems of the form (4.8) in settings where $g(\boldsymbol{\beta})$ has a special structure; a well-known example is LARS (Efron et al., 2004), which can be used to efficiently solve LASSO-type problems in the case where $g(\boldsymbol{\beta})$ is replaced by a least squares objective function. Recently, Combettes and Wajs (2005, Proposition 3.1; Theorem 3.4) proposed a very general class of fixed point algorithms for minimizing $f_1(h)+f_2(h)$, where $f_i(\cdot), i = 1, 2$ are each convex and h takes values in some real Hilbert space $\mathcal{H}$. Hale et al. (2008,

Theorem 4.5) specialize the results of Combettes and Wajs (2005) to the case where $\mathcal{H}$ is some subset of $\mathbb{R}^p$ and $f_2(h) = \sum_{j=1}^{p} |h_i|$. The collective application of these results to the problem of minimizing (4.8) generates an iterated soft-thresholding procedure with an appealingly simple structure. Theorem 4.2.2, given below, states the algorithm, along with conditions under which the algorithm is guaranteed to converge; a proof is provided in Section 4.5. The resulting class of procedures, that is, MM algorithms in which the minimization of (4.8) is carried out via this iterated soft-thresholding procedure, is hereafter referred to as MIST, an acronym for (M)inimization by (I)terated (S)oft (T)hresholding. Two important and useful features of MIST include the absence of high-dimensional matrix inversion and the ability to update each parameter separately.

Theorem 4.2.2. *Suppose the conditions of Theorem 4.1.1 hold. Let* $m(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2$ *be strictly convex with a Lipschitz continuous derivative of order* $L^{-1} > 0$ *for each bounded* $\boldsymbol{\alpha}$*. Then, for any such* $\boldsymbol{\alpha}$ *and a constant* $\varpi \in (0, 2L)$*, the unique minimizer of* (4.8) *can be obtained in a finite number of iterations using iterated soft-thresholding:*

1. *Set* $n = 1$ *and initialize* $\mathbf{b}^{(0)}$
2. *Compute* $\mathbf{d}^{(n)} = \mathbf{b}^{(n-1)} - \varpi\nabla m(\mathbf{b}^{(n-1)})$
3. *Compute* $\mathbf{b}^{(n)} = S(\mathbf{d}^{(n)}; \varpi\boldsymbol{\tau})$*, where for any vectors* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$,

$$S(\mathbf{u}; \mathbf{v}) = \sum_{j=1}^{p} s(u_j, v_j)\, \mathbf{e}_j, \tag{4.9}$$

$\boldsymbol{e}_j$ denotes the j^{th} unit vector for $\mathbb{R}^p$,

$$s(u_j, v_j) = \textit{sign}(u_j)(|u_j| - v_j)_+, \tag{4.10}$$

is the univariate soft-thresholding operator, and

$$\boldsymbol{\tau} = (\tilde{p}'(|\alpha_1|; \boldsymbol{\lambda}_1), \ldots, \tilde{p}'(|\alpha_p|; \boldsymbol{\lambda}_p))^T.$$

4. Stop if converged; else, set $n = n + 1$ *and return to Step 2.*

Note that Theorem 3.4 of Combettes and Wajs (2005) shows that the update in Step 3 can be easily generalized to

$$\mathbf{b}^{(n)} = \mathbf{b}^{(n-1)} + \delta_n \left[S\left(\mathbf{b}^{(n-1)} - \varpi_n \nabla m(\mathbf{b}^{(n-1)}); \varpi_n \boldsymbol{\tau} \right) - \mathbf{b}^{(n-1)} \right],$$

where, for every n, $\varpi_n \in (0, 2L)$ and $\delta_n \in (0, 1]$ are suitably selected sequences of relaxation constants.

Theorem 4.2.2 imposes the relatively strong condition that the gradient of $m(\boldsymbol{\beta})$ is L^{-1}-Lipschitz continuous. The role of this condition, also imposed in Combettes and Wajs (2005, Prop. 3.1; Thm. 3.4), is to ensure that the update at each step of the proposed algorithm is a contraction, thereby guaranteeing its convergence to a fixed point. To see this, note that the update from $\mathbf{b}^{(n)}$ to $\mathbf{b}^{(n+1)}$ in the algorithm involves the mapping $S\big(\mathbf{b} - \varpi \nabla m(\mathbf{b}); \varpi \boldsymbol{\tau}\big)$. For any bounded $\mathbf{b}$ and $\mathbf{a}$, it is easily shown that

$$\| S\big(\mathbf{b} - \varpi \nabla m(\mathbf{b}); \varpi \boldsymbol{\tau}\big) - S\big(\mathbf{a} - \varpi \nabla m(\mathbf{a}); \varpi \boldsymbol{\tau}\big) \| \leq \| \mathbf{b} - \mathbf{a} - \varpi \left(\nabla m(\mathbf{b}) - \nabla m(\mathbf{a}) \right) \|.$$

When $\nabla m(\mathbf{b}) = \nabla m(\mathbf{a})$, the right-hand side reduces to $\|\mathbf{b} - \mathbf{a}\|$, and the resulting mapping is only nonexpansive (not necessarily contractive). However, under strict convexity, this situation can occur only if $\mathbf{b} = \mathbf{a}$. In particular, suppose that $\mathbf{b}^{(n)} \neq \mathbf{b}^{(n-1)}$; then, $\nabla m(\mathbf{b}^{(n)}) \neq \nabla m(\mathbf{b}^{(n-1)})$ and, using the mean value theorem,

$$\begin{aligned} \|\mathbf{b}^{(n+1)} &- \mathbf{b}^{(n)}\| \\ &= \| S\Big(\mathbf{b}^{(n)} - \varpi \nabla m(\mathbf{b}^{(n)}); \varpi \boldsymbol{\tau}\Big) - S\Big(\mathbf{b}^{(n-1)} - \varpi \nabla m(\mathbf{b}^{(n-1)}); \varpi \boldsymbol{\tau}\Big) \| \\ &\leq \| I - \varpi H(\mathbf{b}^{(n)}, \mathbf{b}^{(n-1)}) \| \, \| \mathbf{b}^{(n)} - \mathbf{b}^{(n-1)} \|, \end{aligned}$$

where $H(\mathbf{b}, \mathbf{a}) = \int_0^1 \nabla m(\mathbf{b} + t(\mathbf{a} - \mathbf{b})) dt$. The assumption that the gradient of $m(\boldsymbol{\beta})$ is L^{-1}-Lipschitz continuous now implies that choosing ϖ as indicated guarantees $\| I - \varpi H(\mathbf{b}^{(n)}, \mathbf{b}^{(n-1)}) \| < 1$, thereby producing a contraction.

In view of the generality of the Contraction Mapping Theorem (e.g., Luenberger and Ye, 2008, Thm. 10.2.1), it is possible to relax the requirement that $\nabla m(\boldsymbol{\beta})$ is globally L^{-1}-Lipschitz continuous provided that one selects a suitable starting point. The relevant extension is summarized in the corollary below; one may prove this result in a manner similar to Theorem 4.5 of Hale et al. (2008).

Corollary 4.2.3. *Let the conditions of Theorem 4.1.1 hold. Suppose $\boldsymbol{\alpha}$ is a bounded vector and assume that $m(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2$ is strictly convex and twice continuously differentiable. Then, for a given bounded $\boldsymbol{\alpha}$, there exists a unique minimizer $\boldsymbol{\beta}^*$. Let Ω be a bounded convex set containing $\boldsymbol{\beta}^*$ and define $\lambda_{max}(\boldsymbol{\beta})$ to be the largest eigenvalue of $\nabla^2 m(\boldsymbol{\beta})$. Then, the algorithm of Theorem 4.2.2 converges to $\boldsymbol{\beta}^*$ in a finite number of iterations provided that $\mathbf{b}^{(0)} \in \Omega$, $\lambda^* = \max_{\boldsymbol{\beta}\in\Omega} \lambda_{max}(\boldsymbol{\beta}) < \infty$, and $\varpi \in (0, 2/\lambda^*)$.*

Some useful insight into the form of the proposed thresholding algorithm can be gained by considering the behavior of the penalty derivative term $\tilde{p}'(r; \boldsymbol{\theta})$. Evidently, (P1) implies that $\tilde{p}'(r; \boldsymbol{\theta})$ decreases from its maximum value towards zero as r moves away from the origin. For some penalties (e.g., SCAD, MCP), this derivative actually becomes zero at some finite value of $r > 0$, resulting in a situation in which $\tau_j = \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j) = 0$ for some j. In such cases, the j^{th} component of the vector $S\Big(\mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)}); \varpi\boldsymbol{\tau}\Big)$ simply reduces to the j^{th} component of the argument vector $\mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)})$. In the extreme case where $\boldsymbol{\tau} = \mathbf{0}$, the proposed update simply becomes $\mathbf{b}^{(n+1)} = \mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)})$, an inexact Newton step in which the inverse hessian matrix is replaced by $\varpi\mathbf{I}_p$, $\mathbf{I}_p$ denoting the $p \times p$ identity matrix, and step-size chosen to ensure that the update remains a contraction. Hence, if each of the components of $\mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)})$ are sufficiently large in magnitude, the proposed algorithm simply takes an inexact Newton step towards the solution; otherwise, one or more components of this Newton-like update are subject to soft-thresholding.

4.2.1 Penalized Estimation for Generalized Linear Models

The combination of Theorems 4.1.1, 4.2.2 and Corollary 4.2.3 lead to a useful and stable class of algorithms with the ability to deal with a wide range of penalized regression problems. In settings where $g(\boldsymbol{\beta})$ is strictly convex and twice continuously differentiable, one can safely assume that $h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$ for all choices of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ provided that $\tilde{p}'(r; \boldsymbol{\theta})$ in (P1) is strictly positive for $r > 0$; important examples of statistical estimation problems here include many commonly used linear and generalized linear regression models, semiparametric Cox regression (Cox, 1972), and smoothed versions of the accelerated failure time regression model (cf. Johnson and Strawderman, 2009). The SCAD and MCP penalizations, as well as other penalties having $\tilde{p}'(r; \boldsymbol{\theta}) \geq 0$ for $r > 0$, can also be used; however, additional care is required. In particular, and as pointed out in an earlier remark, if one sets $h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$ for all $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ then convergence of the resulting algorithm to a stationary point is no longer guaranteed by the above results due to the resulting failure of these penalties to induce strict majorization.

The need to use an iterative algorithm for repeatedly minimizing (4.8) is not unusual for the class of MM algorithms. However, it turns out that for certain choices of $g(\boldsymbol{\beta})$, a suitable choice of $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in Theorem 4.2.2 guarantees both strict majorization and permits one to minimize (4.8) in a single iteration, resulting in a single soft-thresholding update at each iteration. Below, it is demonstrated how the MIST algorithm simplifies in settings where $g(\boldsymbol{\beta})$ corresponds to the negative loglikelihood function of a canonically parameterized generalized linear regression model having a bounded hessian function. The result applies to all penalties satisfying condition (P1), including SCAD and MCP. A proof is provided in Section 4.5.

Theorem 4.2.4. *Let $\mathbf{y}$ be $N \times 1$ and suppose the probability distribution of $\mathbf{y}$ follows a generalized linear model with a canonical link and linear predictor $\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}$, where $\widetilde{\boldsymbol{X}} =$*

$[\mathbf{1}_N, \boldsymbol{X}]$ *is* $N \times (p+1)$ *and* $\widetilde{\boldsymbol{\beta}} = [\beta_0, \boldsymbol{\beta}^T]^T$ *is* $(p+1) \times 1$ *with* β_0 *denoting an intercept. Assume that* $g(\widetilde{\boldsymbol{\beta}}) = -\ell(\widetilde{\boldsymbol{\beta}})$*, where*

$$\ell(\widetilde{\boldsymbol{\beta}}) = \sum_{i=1}^{N} [y_i \tilde{\eta}_i - b(\tilde{\eta}_i) + c(y_i)]$$

is the corresponding loglikelihood, $\widetilde{\boldsymbol{\eta}} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}$ *with elements* $\tilde{\eta}_i$ *and* $E[y_i] = b'(\tilde{\eta}_i)$*,* $i = 1, \ldots, N$*, for* $b(\cdot)$ *strictly convex and twice continuously differentiable. Let the penalty function be defined as in* (4.1) *and satisfy (P1); note* β_0 *remains unpenalized. Define*

$$h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = \ell(\widetilde{\boldsymbol{\beta}}) - \ell(\widetilde{\boldsymbol{\alpha}}) - \nabla\ell(\widetilde{\boldsymbol{\alpha}})^T(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\alpha}}) + \varpi^{-1}(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\alpha}})^T(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\alpha}}); \qquad (4.11)$$

where $\tilde{\boldsymbol{\alpha}} \equiv [\alpha_0, \boldsymbol{\alpha}^T]^T$ *is* $(p+1) \times 1$*, and* ϖ *is defined as in Corollary 4.2.3. Then:*

1. *The objective function* (2.3)*, say* $\xi_{glm}(\tilde{\boldsymbol{\beta}})$*, is majorized by*

$$\begin{aligned} \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}) &= -\ell(\tilde{\boldsymbol{\alpha}}) - \nabla\ell(\tilde{\boldsymbol{\alpha}})^T(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\alpha}}) \qquad (4.12) \\ &\quad + \varpi^{-1}(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\alpha}})^T(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\alpha}}) + \sum_{j=1}^{p} (\tau_j |\beta_j| + \gamma_j + \lambda\varepsilon\beta_j^2) \end{aligned}$$

   where $\tau_j = \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)$ *and* $\gamma_j = \tilde{p}(|\alpha_j|; \boldsymbol{\lambda}_j) - \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\alpha_j|$ *are bounded, nonnegative, and functionally independent of* $\tilde{\boldsymbol{\beta}}$*.*

2. *The functions* $g(\widetilde{\boldsymbol{\beta}}) = -\ell(\tilde{\boldsymbol{\beta}})$ *and* $h(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$ *satisfy the regularity conditions of Theorems 4.1.1 and 4.2.2; hence, the corresponding MM algorithm converges to a stationary point of* (2.3)*.*

3. *For each bounded* $\tilde{\boldsymbol{\alpha}}$*,*

   (a) *the minimizer* $\tilde{\boldsymbol{\beta}}^*$ *of* $\xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}})$ *is unique and satisfies*

$$\begin{aligned} \boldsymbol{\beta}^* &= \frac{1}{1 + \varpi\lambda\varepsilon} S\left(\boldsymbol{\alpha} + \frac{\varpi}{2}[\nabla\ell(\tilde{\boldsymbol{\alpha}})]_{\mathcal{A}}, \frac{\varpi}{2}\boldsymbol{\tau}\right), \\ \beta_0^* &= \alpha_0 + \frac{\varpi}{2}[\nabla\ell(\widetilde{\boldsymbol{\alpha}})]_0 \qquad (4.13) \end{aligned}$$

   where $S(\cdot;\cdot)$ *is the soft-thresholding operator defined in* (4.9)*,* $[\cdot]_I$ *denotes the subvector of* $[\cdot]$ *corresponding to index set* I*, and* $\mathcal{A} = \{1, \ldots, p\}$*.*

(b) for each $\tilde{\boldsymbol{\kappa}} \equiv [\kappa_0, \boldsymbol{\kappa}^T]^T \in \mathcal{R}^{(p+1)}$,

$$\xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}^* + \tilde{\boldsymbol{\kappa}}, \tilde{\boldsymbol{\alpha}}) \geq \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\alpha}}) + \varpi^{-1} \|\tilde{\boldsymbol{\kappa}}\|^2 . \tag{4.14}$$

In view of previous results, the result in # 3 of Theorem 4.2.4 shows that the resulting MM algorithm takes a very simple form: given the current iterate $\widetilde{\boldsymbol{\beta}}^{(k)}$,

1. update the unpenalized intercept $\beta_0^{(k)}$:

$$\beta_0^{(k+1)} = \beta_0^{(k)} + \frac{\varpi}{2} \left[\nabla \ell(\tilde{\boldsymbol{\beta}}^{(k)}) \right]_0$$

2. update the remaining parameters $\boldsymbol{\beta}^{(k)}$:

$$\boldsymbol{\beta}^{(k+1)} = \frac{1}{1 + \varpi\lambda\varepsilon} S\left(\boldsymbol{\beta}^{(k)} + \frac{\varpi}{2} [\nabla \ell(\tilde{\boldsymbol{\beta}}^{(k)})]_{\mathcal{A}}; \frac{\varpi}{2} \boldsymbol{\tau}^{(k)} \right), \tag{4.15}$$

where $\boldsymbol{\tau}^{(k)} = (\tilde{p}'(|\beta_1^{(k)}|; \boldsymbol{\lambda}_1), \ldots, \tilde{p}'(|\beta_p^{(k)}|; \boldsymbol{\lambda}_p))^T$.

The specific choice of function $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ clearly serves two useful purposes: (i) it leads to componentwise-soft thresholding; and, (ii) it leads to strict majorization, as is required in condition (iii) of Theorem 4.1.1, allowing one to establish the convergence of MIST for SCAD and other penalties having $\tilde{p}'(r, \boldsymbol{\theta}) = 0$ at some finite $r > 0$.

Evidently, the algorithm above immediately covers the setting of penalized linear regression. For example, suppose that **y** has been centered to remove β_0 from consideration and that the problem has also been rescaled so that **X**, which is now $N \times p$, satisfies the indicated conditions. Then, the results of the Theorem 4.2.4 apply directly with

$$-\ell(\boldsymbol{\beta}) = \frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2 , \ \nabla \ell(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$
$$h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{\varpi} \|\boldsymbol{\beta} - \boldsymbol{\alpha}\|^2 - \frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\alpha}\|^2,$$

where ϖ is defined as in Corollary 4.2.3. For the class of adaptive elastic net penalties (i.e., $\tilde{p}(r; \boldsymbol{\lambda}_j) = \lambda\omega_j r$ in (4.1)), the resulting iterative scheme is exactly that proposed

in De Mol et al. (2008, pg. 17), specialized to the setting of a Euclidean parameter. In particular, $\tau_j = \lambda\omega_j$ and $\gamma_j = 0$ in Theorem 4.2.4, and the proposed update reduces to

$$\boldsymbol{\beta}^{(k+1)} = \frac{1}{\nu + 2\lambda\varepsilon} S\left((\nu\mathbf{I} - \mathbf{X}'\mathbf{X})\,\boldsymbol{\beta}^{(k)} + \mathbf{X}'\mathbf{y}; \lambda \right),$$

where $\nu = 2\varpi^{-1}$. Setting $\nu = 1$ and $\varepsilon = 0$ yields the iterative procedure proposed in Daubechies et al. (2004), provided that $\mathbf{X}'\mathbf{X}$ is scaled such that $\mathbf{I} - \mathbf{X}'\mathbf{X}$ is positive definite. The proposed MIST algorithm extends these iterative componentwise soft-thresholding procedures to a much wider class of penalty and data fidelity functions.

In an interesting recent but unpublished paper, Mazumder et al. (2009) propose the SparseNet algorithm, a coordinatewise descent algorithm for minimizing objective functions of the form (2.3) with $g(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2$, $\varepsilon = 0$ and $p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ a family of penalty functions satisfying (4.1) and several additional regularity conditions. Their specification includes the LASSO, SCAD and MCP penalties, as well as several other examples of nonconvex penalties. The full SparseNet algorithm intends to generate the solution surface as a function of the penalty parameter λ and a parameter γ indexing the penalty family (i.e., restricted to a two dimensional grid). While the algorithm incorporates a number of useful features, solutions are found for each (λ, γ) pair using a simple coordinate descent algorithm. In the case of the LASSO penalty ($\gamma = \infty$) and provided $\mathbf{X}$ is column-standardized, this coordinate descent algorithm is almost identical to the componentwise soft-thresholding algorithm proposed in Daubechies et al. (2004) (hence MIST), the primary differences stemming from the form of the iterative update (i.e., the use of a simultaneous update implemented via componentwise soft-thresholding versus cyclical application of the soft-thresholding operator). For other penalties, such as SCAD and MCP, the coordinatewise updates utilized by SparseNet rely on so-called generalized thresholding operators (cf. She, 2009), departing more substantially from the iterated soft-thresholding procedure used in the MIST algorithm. Mazumder et al. (2009) provide an explicit proof of the convergence of the solution sequence obtained

for a given (λ, γ) pair. The regularity conditions under which these results are obtained appear to be similarly weak to those required by Theorem 4.2.4 (i.e., applied to the penalized least squares problem). However, unlike MIST, it not obvious how to extend the SparseNet algorithm to more general choices of $g(\boldsymbol{\beta})$ in the absence of reparameterizations that permit componentwise separation of parameters.

The restriction to canonical generalized linear models in Theorem 4.2.4 is imposed to ensure strict convexity of the negative loglikelihood. Our results are easily modified to handle non-canonical generalized linear models, provided the negative loglikelihood remains strictly convex in $\widetilde{\boldsymbol{\beta}}$ and the hessian can be appropriately bounded. Interestingly, not all canonically parameterized generalized linear models satisfy the regularity conditions of Theorem 4.2.4. One such important class of problems is penalized likelihood estimation for Poisson regression models. For example, in the classical setting of N independent Poisson observations with $E[Y_i|\widetilde{\mathbf{X}}_i] = d_i \exp\{\tilde{\mathbf{x}}_i^T \widetilde{\boldsymbol{\beta}}\}$ for known constants $d_1 \ldots d_N$, we have (up to irrelevant constants) $\ell(\widetilde{\boldsymbol{\beta}}) = -\sum_{i=1}^N f_i(\tilde{\mathbf{x}}_i^T \widetilde{\boldsymbol{\beta}})$, where

$$f_i(u) = d_i e^u - y_i u.$$

It is easy to see that $\nabla \ell(\widetilde{\boldsymbol{\beta}})$, hence $\nabla m(\widetilde{\boldsymbol{\beta}})$, is locally but not globally Lipschitz continuous; hence, it is not possible to choose a matrix $\mathbf{C} = \varpi^{-1}\mathbf{I}$ such that (4.12) everywhere majorizes $\xi_{glm}(\widetilde{\boldsymbol{\beta}})$. Nevertheless, progress remains possible. For example, Corollary 4.2.3 implies that that one can still use a single update of the form (4.15) provided that a suitable Ω, hence $\mathbf{C}$ and $\widetilde{\boldsymbol{\beta}}^{(0)}$, can be identified. Alternatively, using results summarized in Becker et al. (1997), one can instead majorize $-\ell(\widetilde{\boldsymbol{\beta}})$ for any bounded $\boldsymbol{\alpha}$ using $k(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = \sum_{j=0}^p k_j(\beta_j; \alpha_j)$, with

$$k_j(\beta_j; \alpha_j) = \sum_{i=1}^n I\{x_{ij} \neq 0\}\, \theta_{ij}\, f_i\left(\frac{x_{ij}}{\theta_{ij}}(\beta_j - \alpha_j) + \tilde{\mathbf{x}}_i^T \widetilde{\boldsymbol{\alpha}}\right),$$

where, for every i, $\theta_{ij} \geq 0$ are any sequence of constants satisfying $\sum_{j=0}^p \theta_{ij} = 1$ and $\theta_{ij} > 0$ if $x_{ij} \neq 0$. Of importance here is the fact that $k_j(\beta_j; \alpha_j)$ is a strictly convex

function of β_j and does not depend on β_k for $k \neq j$. One may now take $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ in Theorem 4.1.1 as being equal to $k(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) + \ell(\widetilde{\boldsymbol{\beta}})$, leading to the minimization of

$$\xi^{SUR}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) \propto \sum_{j=1}^{p} [k_j(\beta_j; \alpha_j) + \lambda\varepsilon\beta_j^2 + \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|] + k_0(\beta_0, \alpha_0). \tag{4.16}$$

In particular, componentwise soft-thresholding is replaced by componentwise minimization of (4.16), the latter being possible using any algorithm capable of minimizing a continuous nonlinear function of one variable.

Remark 4.2.5. *The Cox proportional hazards model (Cox, 1972), while not a generalized linear model, shares the essential features of the generalized linear model utilized in Theorem 4.2.4. In particular, the negative log partial likelihood, say $g(\boldsymbol{\beta}) = -\ell_p(\boldsymbol{\beta})$, is strictly convex, twice continuously differentiable, and has a bounded hessian (e.g., Bohning and Lindsay, 1988; Andersen et al., 1993). Consequently, Theorem 4.2.4 and its proof are easily modified for this setting upon taking $g(\boldsymbol{\beta})$ as indicated, setting $h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \ell_p(\boldsymbol{\beta}) - \ell_p(\boldsymbol{\alpha}) - \nabla\ell_p(\boldsymbol{\alpha})^T(\boldsymbol{\beta} - \boldsymbol{\alpha}) + \varpi^{-1}\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|^2$, and taking ϖ as defined as in Corollary 4.2.3.*

4.2.2 Accelerating Convergence

Similarly to the EM algorithm, the stability and simplicity of the MM algorithm frequently comes at the price of a slow convergence rate. Numerous methods of accelerating the EM algorithm have been proposed in the literature; see McLachlan and Krishnan (2008) for a review. Recently, Varadhan and Roland (2008) proposed a new method for EM called SQUAREM, obtained by "squaring" an iterative Steffensen-type (STEM) acceleration method. Both STEM and SQUAREM are structured for use with iterative mappings of the form $\theta^{(k+1)} = M(\theta^{(k)})$, $k = 0, 1, 2, \ldots$, hence applicable to both EM

and MM algorithms. Specifically, the acceleration update for SQUAREM is given by

$$\begin{aligned}\theta^{(k+1)} &= \theta^{k} - 2\gamma_k(M(\theta^{(k)}) - \theta^{(k)}) + \gamma_k^2[M(M(\theta^{(k)})) - 2M(\theta^{(k)}) + \theta^{(k)}] \\ &= \theta^{(k)} - 2\gamma_k r_k + \gamma_k^2 v_k, \end{aligned} \tag{4.17}$$

where $r_k = M(\theta^{(k)}) - \theta^{(k)}$ and $v_k = (M(M(\theta^{(k)})) - M(\theta^{(k)})) - r_k$ for an adaptive steplength γ_k. Varadhan and Roland (2008) suggest several steplength options, with preference for choice $\gamma_k = -\|r_k\|/\|v_k\|$. Roland and Varadhan (2005) provide a proof of local convergence for SQUAREM under restrictive conditions on the EM mapping $M(\theta)$, while Varadhan and Roland (2008) outline a proof for global convergence for versions of SQUAREM that employ a back-tracking strategy. The effectiveness of SQUAREM applied to the simplified form of the MIST algorithm, hereafter denoted SQUAREM^2, is examined in Section 4.3.3.

4.3 Simulation Results

The simulation results summarized below are intended to compare the estimates of $\boldsymbol{\beta}$ obtained from existing methods to those obtained using the simplified MIST algorithm of Theorem 4.2.4. In particular, the performance of MIST for the class of penalized linear and generalized linear models is considered, demonstrating its capability of recovering the solutions provided by existing algorithms when both algorithms are forced to use the same set of "tuning" parameters (i.e., penalty parameter(s), plus any additional parameters required to define the penalty itself). In cases where multiple local minima can arise, it is further shown that the MIST algorithm often tends to find solutions with lower objective function evaluations in comparison with existing algorithms, provided these algorithms utilize the same choice of starting value.

4.3.1 Example 1: Linear Model

Let $\mathbf{1}_m$ and $\mathbf{0}_m$ respectively denote m-dimensional vectors of ones and zeros. Then, following Zou and Zhang (2009), the data was generated according to the linear regression model

$$y = \mathbf{x}'\boldsymbol{\beta}^* + \epsilon \tag{4.18}$$

where $\boldsymbol{\beta}^* = (3 \cdot \mathbf{1}_q^T, \mathbf{0}_{p-q}^T)^T$ is a p-dimensional vector with intrinsic dimension $q = 3[p/9]$, $\epsilon \sim N(0, \sigma^2)$, and $\mathbf{x}$ follows a p-dimensional multivariate normal distribution with zero mean and covariance matrix Σ having elements $\Sigma_{j,k} = \rho^{|j-k|}$, $1 \leq k, j \leq p$. We considered $\sigma \in \{1, 3\}$, $\rho \in \{0.0, 0.5, 0.75\}$ and $p \in \{35, 81\}$ for $N = 100$ independent observations.

Penalized least squares estimation is considered for five popular choices of penalty functions, all of which are currently implemented in the R software language (R Development Core Team, 2005): LAS, ALAS, EN, AEN, and SCAD. The LAS, ALAS, EN and AEN penalties are all convex and lead to unique solutions under mild conditions; the SCAD penalty is concave and the resulting minimization problem may have multiple solutions. In each case, we used existing software for computing solutions subject to these penalizations and compared those results to the solutions computed using the MIST algorithm.

Regarding existing methods, we respectively used the *lars* (Hastie and Efron, 2007) and *elasticnet* (Zou and Hastie, 2008) packages for computing solutions in the case of the LAS and EN penalties. For the ALAS and AEN penalties, we used software kindly provided by Zou and Zhang (2009) which makes use of the *elasticnet* package. The weights for the AEN penalty are computed using $\omega_j = |\hat{\beta}_j^{EN}|^{-\gamma}$, $j = 1, \ldots, p$, where $\hat{\boldsymbol{\beta}}^{EN}$ is an EN estimator and γ is a positive constant. Using EN-based weights in the AEN fitting algorithm necessitates tuning parameter specification for both EN and AEN.

As in Zou and Zhang (2009), the L_1 parameters λ (λ_1 in their notation) are allowed to differ, whereas the L_2 parameters ε (λ_2 in their notation) are forced to be the same. Evidently, setting $\varepsilon = 0$ ($\lambda_2 = 0$) results in the ALAS solution. For the SCAD penalty, we considered the estimator of Kim et al. (2008) (HD), as well the one-step (1S) and LLA estimators of Zou and Li (2008). The code for the first two methods was kindly provided by their respective authors; the LLA estimator was computed using the *SIS* package (Fan et al., 2009a). Choice $a = 3.7$ was used for all implementations of SCAD.

We considered finding solutions using penalty parameters in the set $\Lambda = \{0.1, 1, 5, 10, 20, 100\}$. In particular, for LAS and SCAD, $\lambda = \lambda_1 \in \Lambda$. For EN, both $\lambda = \lambda_1 \in \Lambda$ and $\lambda\varepsilon = \lambda_2 \in \Lambda$. For simplicity, we fixed the weights for AEN for a given λ_2 by selecting the 'best' $\hat{\boldsymbol{\beta}}^{EN}$ among the six estimators involving $\lambda = \lambda_1 \in \Lambda$ based on a BIC-like criteria. Likewise for ALAS, the weights were computing using the 'best' $\hat{\boldsymbol{\beta}}^{LAS}$ among the six estimators involving $\lambda = \lambda_1 \in \Lambda$. The parameter γ for the ALAS and AEN penalties was respectively set to three and five for $p = 35$ and $p = 81$.

For the strictly convex objective functions associated with the LAS, ALAS, EN, and AEN penalties, we simply used a starting value of $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p$. For SCAD, three different starting values for the MIST, HD, and LLA SCAD algorithms were considered: $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p$, $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}_{ml}$ (i.e., the unpenalized least squares estimate), and $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}_{1S,\lambda}$ (i.e., the one-step estimate computed using the penalty λ). As in Zou and Li (2008), the one-step estimator 1S is computed using $\widehat{\boldsymbol{\beta}}_{ml}$, an appropriate choice when $N > p$.

The convergence criteria used by the existing software packages were used without alteration in this simulation study. The convergence criteria used for MIST were as follows: the algorithm stopped if either (i) the normed difference of successive iterates was less than 10^{-6} (convergence of coefficients); or, (ii) the difference of the objective function evaluated at successive iterates was less than 10^{-6} and the number of iterations

Table 4.1: Maximum average normed differences ($\times 10^5$) over $B = 100$ simulations for Examples 1 (LM) and 2 (GLM).

	$LM : \sigma = 1$			$LM : \sigma = 3$			GLM		
ρ	0	0.5	0.75	0	0.5	0.75	0	0.5	0.75
$p = 35$							$q = 25$		
LAS	0.10	0.35	1.45	0.10	0.37	1.56	0.07	4.28	6.17
ALAS	0.03	0.14	0.64	0.05	0.21	1.00	1.84	2.86	3.76
EN	0.07	0.19	0.50	0.07	0.20	0.51	2.30	5.61	8.68
AEN	0.03	0.10	0.33	0.04	0.13	0.36	1.47	3.35	5.27
$p = 81$							$q = 75$		
LAS	1.73	3.82	11.76	2.33	5.78	18.99	0.10	6.97	9.94
ALAS	0.12	0.38	1.58	0.35	1.03	4.39	1.34	2.55	3.30
EN	0.31	0.49	0.87	0.31	0.49	0.88	2.35	4.64	6.56
AEN	0.14	0.22	0.56	0.16	0.26	0.56	1.27	2.29	2.85

exceeded 10^6 (convergence of optimization). Due to the large number of comparisons and highly intensive nature of the computations, we ran $B = 100$ simulations for each choice of ρ, σ, and p. We report the results for the convex penalties in Table 4.1 and those for the SCAD penalty in Tables 4.2 and 4.3.

In Table 4.1, we summarize the average normed difference between the solution obtained using existing software and that obtained using MIST, $\left\| \hat{\boldsymbol{\beta}}_{exist} - \hat{\boldsymbol{\beta}}_{mist} \right\|$, over the $B = 100$ simulations; in particular, we report in the two leftmost panels the maximum value of this difference, computed across all combinations of tuning parameters. These maximum differences (all of which are multiplied by 10^5) are remarkably small for all (A)LAS and (A)EN penalties, indicating that MIST recovers the same (unique) solutions as the existing algorithms. Interestingly, the values for LAS are slightly larger than the rest, where the maximum differences all resulted from the smallest value of λ considered ($\lambda = 0.1$). In these cases, the algorithm tended to stop using the objective function criteria rather than the (stricter) coefficient criteria, resulting in slightly larger differences on average.

The results for SCAD are reported in Tables 4.2 ($p = 35$) and 4.3 ($p = 81$) and display (i) the average normed differences, multiplied by 10^3, for each combination of λ, ρ, σ, p and starting value; and, (ii) the proportion of simulated datasets in which the MIST solution yields a lower evaluation of the objective function in comparison with the solution obtained using another method for the indicated choice of starting value. We remark here that SCAD penalties used in the existing implementations are multiplied by a factor of N, i.e., $p(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \sum_{j=1}^{p} N\tilde{p}_S(|\beta_j|; \lambda, a)$, so the MIST implementation incorporates this factor of N as well. The results for $\lambda = 100$ are not shown, as the solution was $\mathbf{0}_p$ in all cases. In comparison with the convex penalties, larger normed differences are observed, even when controlling for the use of the same starting value. Such differences are a result of two important features of the SCAD optimization problem: (i) the possible existence of several local minima; and, (ii) the fact that the MIST, HD, and LLA algorithms each take a different path from a given starting value towards one of these solutions. For example, while each of the LLA, MIST, and HD algorithms involve majorization of the objective function using a LASSO-type surrogate objective function, both the majorization and minimization of the resulting surrogate function are carried out differently in each case. In particular, the LLA algorithm, as implemented in *SIS*, majorizes only the penalty term and adapts the LASSO code in *glmpath* in order to minimize the corresponding surrogate objective function at each step. The HD algorithm is similar in spirit, but instead decomposes the penalty term into a sum of a concave and convex function and utilizes the algorithm of Rosset and Zhu (2007) to minimize the corresponding surrogate objective function. The MIST algorithm instead uses the same penalty majorization as the LLA algorithm, but additionally majorizes the negative loglikelihood term in a way that permits minimization of the surrogate function in a single soft-thresholding step. Each procedure therefore takes a different path towards a solution, even when given the same starting value.

We remark here that differences must also expected between any of LLA, HD, MIST and the one-step solution 1S; from an optimization perspective, the one-step estimate is the result of running just one iteration of the LLA algorithm, starting from the unpenalized least squares estimator $\widehat{\boldsymbol{\beta}}_{ml}$ (Zou and Li, 2008), and only provides an approximation the solution to the desired minimization problem. All other methods (LLA, MIST, HD) iterate until some local minimizer (or stationary point) is reached. For example, when using either $\widehat{\boldsymbol{\beta}}_{ml}$ or $\widehat{\boldsymbol{\beta}}_{1S,\lambda}$ as the starting value, MIST always found a solution that produced a lower evaluation of the objective function in comparison to $\widehat{\boldsymbol{\beta}}_{1S,\lambda}$. However, when using the null starting value of $\mathbf{0}_p$, the one-step estimator did occasionally result in a lower objective function evaluation in cases involving smaller values of λ. This behavior is not terribly surprising; with small λ, the one-step solution should generally be close to the unpenalized least squares solution, as the objective function itself is likely to be dominated by the least squares term.

Of all the SCAD algorithms considered here, MIST and LLA tended to find the most similar solutions (i.e., have the smallest normed differences). For the cases in which the LLA solution had lower objective function evaluations, all of the MIST solutions were also LLA solutions; i.e, when starting the LLA algorithm with the MIST solution, the algorithm terminated at the starting value (i.e., the LLA solution coincides with the MIST solution). With the exception of three of these cases, starting the MIST algorithm with the LLA solution also resulted in the same behavior. For the most part, the HD and MIST algorithms also gave similar results, with one source of difference being the respective stopping criteria used. The stopping criteria for HD, assessed in order, are as follows: (1) 'convergence of optimization': stop if the absolute value of the difference of the objective evaluated at successive iterates is less than 1e-6; (2) 'convergence of penalty gradient': stop if the sum of the absolute value of the differences of the derivative of the centered penalty evaluated at successive iterates is less than 1e-6, (3) 'convergence

of coefficients:' stop if the sum of the absolute value of the differences of successive iterates is less than 1e-6, and (4) 'jump-over' criteria: stop if the objective at the previous iterate plus 1e-6 was less than the objective at the current iterate. After careful analysis of the results, we can assert the following:

- The MIST solution usually has the same or a lower evaluation of the objective function in comparison with HD, regardless of starting value.
- HD tends to have the greatest difficulty in cases of high correlation between predictors, a likely result of the fact that this algorithm relies on the variance of the unpenalized least squares estimator, hence matrix inversion, to take steps towards solution. In contrast, MIST requires no matrix inversion.

On balance, the MIST algorithm performs as well or better than LLA and HD in locating minimizers in nearly all cases. As suggested above, variation in the solutions found can be traced to the path each algorithm takes towards a solution and differences in stopping criteria. Remarkably, in cases when the correlation among predictors was low, the choice of starting value made little difference for MIST; either the same solution was found for all starting values or none of the starting values dominated in terms of finding the lower or equivalent objective evaluations. In settings involving higher correlation, however, using either $\mathbf{0}_p$ or the $1S$ starting values tended to result in the lower evaluations of the objective function in comparison with using the unpenalized least squares solution. Similar behavior was observed for the LLA algorithm. In contrast, the choice of starting value had a much larger impact on the performance of the HD estimator; in particular, the use of $\mathbf{0}_p$ as a starting value typically resulted in the lowest objective function evaluations when compared to using a non-null starting value.

Table 4.2: Algorithm performance in Example 1 (LM: $p = 35, N = 100$) for SCAD penalty. The column 'avg' is the average normed differences $\times 10^3$ between the MIST solution and the existing method's solution; 'prop' is the proportion of MIST solutions whose objective function evaluation was less than or equal to that of the existing method's solution.

		$\sigma = 1$						$\sigma = 3$					
$\boldsymbol{\beta}^{(0)}$		$\mathbf{0}_p$		$\widehat{\boldsymbol{\beta}}_{ml}$		$\widehat{\boldsymbol{\beta}}_{1S,\lambda}$		$\mathbf{0}_p$		$\widehat{\boldsymbol{\beta}}_{ml}$		$\widehat{\boldsymbol{\beta}}_{1S,\lambda}$	
ρ	method	avg	prop	avg	prop	avg	prop	avg	prop	avg	prop	avg	prop
$\lambda = .1$													
0	HD	15.71	1.00	15.41	1.00	17.93	1.00	468.55	1.00	2076.40	1.00	55.17	1.00
	1S	99.13	1.00	99.13	1.00	99.13	1.00	211.17	1.00	211.16	1.00	211.16	1.00
	LLA	0.43	1.00	0.46	1.00	0.46	1.00	2.07	1.00	1.96	1.00	2.02	1.00
0.5	HD	7.07	0.99	10.72	1.00	2.04	1.00	269.85	0.97	218.94	0.94	130.76	0.98
	1S	192.22	1.00	192.01	1.00	192.00	1.00	483.89	0.98	421.17	1.00	419.15	1.00
	LLA	6.65	0.99	0.62	1.00	0.60	1.00	57.87	0.96	12.84	0.99	2.37	1.00
0.75	HD	29.25	0.99	105.39	0.92	66.83	0.96	2335.23	1.00	2758.43	0.98	2731.10	0.99
	1S	575.09	1.00	488.09	1.00	486.19	1.00	1417.97	0.86	604.26	1.00	629.21	1.00
	LLA	23.81	0.98	23.34	0.99	1.67	0.99	558.56	0.73	69.30	0.96	44.87	0.98
$\lambda = 1$													
0	HD	6.22	1.00	22.87	1.00	19.99	1.00	9.44	1.00	35.16	1.00	14.65	1.00
	1S	694.59	1.00	694.57	1.00	694.57	1.00	844.68	1.00	844.67	1.00	844.67	1.00
	LLA	1.64	1.00	1.71	1.00	1.74	1.00	1.47	1.00	1.47	1.00	1.43	1.00
0.5	HD	300.62	0.98	34.09	1.00	115.76	0.98	303.98	0.96	140.26	1.00	94.90	1.00
	1S	4489.01	1.00	4276.77	1.00	4261.64	1.00	3547.69	1.00	3254.16	1.00	3254.16	1.00
	LLA	296.53	0.98	7.10	1.00	88.14	0.98	248.82	0.96	2.66	1.00	2.66	1.00
0.75	HD	3083.00	0.68	1980.40	0.89	1138.53	0.96	1476.59	0.84	1669.60	0.93	868.21	0.97
	1S	7224.77	1.00	5491.09	1.00	5622.21	1.00	5682.04	0.96	3835.30	1.00	3748.35	1.00
	LLA	2802.66	0.66	1121.80	0.85	293.50	0.96	1365.76	0.83	918.63	0.89	433.66	0.96
$\lambda = 5$													
0	HD	18.18	1.00	18.18	1.00	18.18	1.00	17.73	1.00	17.73	1.00	17.73	1.00
	1S	48.23	1.00	48.23	1.00	48.23	1.00	63.63	1.00	63.63	1.00	63.63	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.5	HD	0.01	1.00	0.01	1.00	0.01	1.00	0.01	1.00	0.01	1.00	0.01	1.00
	1S	3696.85	1.00	3696.85	1.00	3696.85	1.00	3751.96	1.00	3751.96	1.00	3751.96	1.00
	LLA	0.02	1.00	0.09	1.00	0.08	1.00	0.03	1.00	0.14	1.00	0.08	1.00
0.75	HD	0.27	1.00	0.27	1.00	98.05	1.00	19.20	0.99	19.21	0.99	99.95	0.99
	1S	3977.93	1.00	3977.93	1.00	4045.81	1.00	4170.49	1.00	4170.49	1.00	4180.79	1.00
	LLA	0.27	1.00	0.45	1.00	98.35	1.00	19.00	0.99	19.20	0.99	100.05	0.99
$\lambda = 10$													
0	HD	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	1S	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.5	HD	57.33	1.00	57.33	1.00	57.33	1.00	53.80	1.00	53.80	1.00	53.80	1.00
	1S	501.86	1.00	501.86	1.00	501.86	1.00	497.87	1.00	497.87	1.00	497.87	1.00
	LLA	0.01	1.00	0.03	1.00	0.01	1.00	0.01	1.00	0.04	1.00	0.01	1.00
0.75	HD	0.41	1.00	0.41	1.00	0.41	1.00	0.53	1.00	0.53	1.00	0.53	1.00
	1S	4206.65	1.00	4206.65	1.00	4206.65	1.00	4261.12	1.00	4261.12	1.00	4261.12	1.00
	LLA	0.09	1.00	0.30	1.00	0.14	1.00	0.07	1.00	0.36	1.00	0.10	1.00
$\lambda = 20$													
0	HD	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	1S	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.5	HD	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	1S	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.75	HD	33.90	1.00	33.90	1.00	33.90	1.00	35.46	1.00	35.46	1.00	35.46	1.00
	1S	47.21	1.00	47.21	1.00	47.21	1.00	46.90	1.00	46.90	1.00	46.90	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.06	1.00	0.00	1.00

4.3.2 Example 2: Binary Logistic Regression

As in Example 1, we considered the LAS, ALAS, EN, AEN, and SCAD penalties. There are at least two R packages that allow penalization using the LAS and EN penal-

Table 4.3: Algorithm performance in Example 1 (LM: $p = 81, N = 100$) for SCAD penalty. The column 'avg' is the average normed differences ($\times 10^3$) between the MIST solution and the existing method's solution; 'prop' is the proportion of MIST solutions whose objective function evaluation was less than or equal to that of the existing method's solution.

$\boldsymbol{\beta}^{(0)}$		$\sigma = 1$						$\sigma = 3$					
		$\mathbf{0}_p$		$\widehat{\boldsymbol{\beta}}_{ml}$		$\widehat{\boldsymbol{\beta}}_{1S,\lambda}$		$\mathbf{0}_p$		$\widehat{\boldsymbol{\beta}}_{ml}$		$\widehat{\boldsymbol{\beta}}_{1S,\lambda}$	
ρ	method	avg	prop	avg	prop	avg	prop	avg	prop	avg	prop	avg	prop
	$\lambda = .1$												
0	HD	828.22	1.00	1211.97	1.00	962.10	1.00	4615.10	1.00	5414.49	1.00	5350.54	1.00
	1S	753.85	1.00	753.84	1.00	753.84	1.00	2836.29	0.90	1314.46	1.00	1366.62	1.00
	LLA	1.60	1.00	1.67	1.00	1.64	1.00	1181.62	0.76	382.17	0.82	223.32	0.94
0.5	HD	5992.88	1.00	6008.14	1.00	5994.86	1.00	8002.08	1.00	9530.30	1.00	9546.21	1.00
	1S	1217.02	1.00	1202.01	1.00	1201.30	1.00	4619.22	0.88	1473.61	1.00	1403.36	1.00
	LLA	24.78	0.97	1.33	1.00	8.50	0.99	2123.22	0.57	576.65	0.83	232.10	0.91
0.75	HD	12018.61	1.00	12042.97	1.00	12042.90	1.00	13582.93	1.00	16580.85	1.00	16569.80	1.00
	1S	2492.18	1.00	2327.76	1.00	2330.54	1.00	8204.45	0.60	1215.98	1.00	1181.16	1.00
	LLA	36.95	0.98	90.89	0.97	90.69	0.96	3517.93	0.50	607.08	0.78	252.75	0.89
	$\lambda = 1$												
0	HD	1421.70	1.00	3595.88	1.00	2296.03	1.00	1552.11	0.98	3258.39	1.00	2231.63	1.00
	1S	7121.11	1.00	6977.35	1.00	6976.16	1.00	7485.99	1.00	7182.76	1.00	7182.76	1.00
	LLA	50.48	0.99	64.69	0.99	4.59	1.00	231.48	0.97	107.36	1.00	140.97	1.00
0.5	HD	4505.31	0.93	6764.71	0.88	4973.51	0.98	4571.62	0.97	6473.05	0.89	6150.70	0.96
	1S	11973.29	1.00	10301.59	1.00	10238.21	1.00	12411.82	1.00	9674.64	1.00	9781.43	1.00
	LLA	1622.24	0.89	661.69	0.95	622.25	0.96	1682.66	0.89	1785.73	0.86	517.91	0.97
0.75	HD	11166.35	0.75	16786.90	0.57	11642.59	0.84	12834.39	0.81	14964.11	0.66	10110.16	0.90
	1S	16953.51	1.00	9125.82	1.00	9225.76	1.00	17174.91	0.99	8828.81	1.00	8549.86	1.00
	LLA	6379.56	0.50	4295.69	0.63	787.30	0.93	6904.11	0.52	3637.68	0.74	812.28	0.94
	$\lambda = 5$												
0	HD	12.35	1.00	12.35	1.00	12.35	1.00	13.00	1.00	13.00	1.00	13.00	1.00
	1S	1072.70	1.00	1072.70	1.00	1072.70	1.00	1114.13	1.00	1114.13	1.00	1114.13	1.00
	LLA	0.01	1.00	0.05	1.00	0.01	1.00	0.01	1.00	0.07	1.00	0.01	1.00
0.5	HD	28.71	1.00	28.71	1.00	28.71	1.00	0.43	1.00	0.42	1.00	0.43	1.00
	1S	6793.73	1.00	6793.73	1.00	6793.73	1.00	6831.01	1.00	6831.01	1.00	6831.01	1.00
	LLA	0.38	1.00	0.54	1.00	0.49	1.00	0.43	1.00	0.58	1.00	0.57	1.00
0.75	HD	4998.08	0.88	4963.08	0.88	4292.65	0.97	5753.61	0.92	5772.76	0.95	5192.19	0.98
	1S	11191.83	1.00	11188.02	1.00	12029.12	1.00	11917.77	1.00	11971.47	1.00	12485.14	1.00
	LLA	1217.39	0.90	1252.65	0.89	1060.08	0.99	861.72	0.95	937.76	0.94	1018.59	0.98
	$\lambda = 10$												
0	HD	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	1S	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.5	HD	6.69	1.00	6.69	1.00	6.69	1.00	5.80	1.00	5.80	1.00	5.80	1.00
	1S	2883.52	1.00	2883.52	1.00	2883.52	1.00	2906.35	1.00	2906.35	1.00	2906.35	1.00
	LLA	0.03	1.00	0.20	1.00	0.03	1.00	0.02	1.00	0.20	1.00	0.02	1.00
0.75	HD	122.19	1.00	122.19	1.00	122.19	1.00	107.93	1.00	107.93	1.00	107.93	1.00
	1S	8835.88	1.00	8835.88	1.00	8835.87	1.00	8874.85	1.00	8874.85	1.00	8874.84	1.00
	LLA	0.08	1.00	0.54	1.00	0.32	1.00	0.10	1.00	0.53	1.00	0.35	1.00
	$\lambda = 20$												
0	HD	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	1S	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.5	HD	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	1S	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
	LLA	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.75	HD	21.76	1.00	21.76	1.00	21.76	1.00	17.70	1.00	17.70	1.00	17.70	1.00
	1S	3997.88	1.00	3997.88	1.00	3997.88	1.00	4014.29	1.00	4014.30	1.00	4014.29	1.00
	LLA	0.05	1.00	0.43	1.00	0.06	1.00	0.07	1.00	0.38	1.00	0.08	1.00

ties: *glmpath* (Park and Hastie, 2007), which handles binomial and poisson regression using a "predictor-corrector" method, and *glmnet* (Friedman et al., 2008), which handles binomial and multinomial regression using cyclical coordinate descent. Both methods

can be tuned to find the same solutions, so for ease of presentation we only consider the results of *glmnet* for comparison in the tables and analysis below. The *SIS* package (Fan et al., 2009a) permits computations with the ALAS, AEN, and SCAD penalties using modifications of the Park and Hastie (2007) code. For SCAD, we compared the results of MIST to the results from the one-step (1S) algorithm (GLM version, Zou and Li, 2008) using the code provided from the authors and the LLA algorithm as implemented in Fan et al. (2009a). In all cases, MIST was implemented to match the scaling used in the existing algorithms.

As before, we only considered comparing solutions that use the same combination of tuning parameters; for the present example, the set considered here is $\Lambda = \{0.001, 0.01, 0.05, 0.1, 0.2, 1.00\}$, reflecting a need to accommodate the different scaling of the problem. The data generation scheme for this example was loosely based on the simulation study found in Friedman et al. (2008). Binary response data were generated according to a logistic (rather than linear) regression model using $p_i = [1 + \exp(-\mathbf{x}_i'\boldsymbol{\beta}^*)]^{-1}$, $i = 1, \ldots, N = 1000$, where $\boldsymbol{\beta}^*$ is a $p-$vector with elements $\beta_j = 3 \times (-1)^j \exp(-2(j-1)/200),\ j = 1, \ldots, q,\ q \in \{25, 75\}$, and the remaining $100 - q$ components set to zero. Here, $\mathbf{x}_i$ follows a p-dimensional multivariate normal distribution with zero mean and covariance $\boldsymbol{\Sigma} = 3^{-2}\mathbf{P}$ where correlation matrix $\mathbf{P}$ is such that each pair of predictors has the same population correlation ρ. We considered three such correlations, $\rho \in \{0.0, 0.5, 0.75\}$.

For the $B = 100$ simulations, the maximum (across different tuning parameters) average normed difference between the existing and proposed methods, multiplied by 10^5, are reported for each of the strictly convex cases in the right-most panel of Table 4.1. As before, these maximums are generally remarkably small, indicating that MIST can recover the same (unique) solutions as the existing algorithms. The results for SCAD

are reported in Table 4.4, which displays the same information as in the corresponding tables from Example 1; the HD comparisons are omitted here as the methodology and code were only developed for the case of penalized least-squares. In the GLM setting, the 1S estimator is computed by applying the LARS (Efron et al., 2004) algorithm to a quadratic approximation of the negative loglikelihood function evaluated at the MLE. Thus, 1S takes 'one step' towards minimizing the objective function; in contrast, both MIST and LLA iterate until a stationary point, usually a local minimizer, is found. As in the linear model case, LLA uses *glmpath* to minimize the surrogate at each step, whereas the MIST algorithm uses a single application of the soft thresholding operator to minimize the surrogate at each step.

In this example, the starting value carried even greater importance in comparison with the linear model setting. In particular, in the case of MIST, the combination of a $\mathbf{0}_p$ starting value and small penalty parameter led to solutions with objective function evaluations that were substantially larger in comparison with those obtained using either $\widehat{\boldsymbol{\beta}}_{ml}$ and $\widehat{\boldsymbol{\beta}}_{1S,\lambda}$. Such behavior may be directly attributed to the fact that the ML and 1S starting values either minimize or nearly minimize the negative loglikelihood portion of the objective function, the dominant term in the objective function when λ is "small." In contrast, a $\mathbf{0}_p$ starting value led to the best minimization performance for "large" λ; upon reflection, this is also not very surprising, since large penalties induce greater sparsity and $\mathbf{0}_p$ is the sparsest possible solution.

There were a few settings in which the 1S estimator resulted in a lower objective function evaluation in comparison with applying MIST started at $\widehat{\boldsymbol{\beta}}_{ml}$. This reflects the fact that several local minima can exist for non-convex penalties like SCAD. In addition, and as was observed before, using the 1S solution as a starting value always led to MIST finding a solution with a lower evaluation of the objective function in comparison with

Table 4.4: Algorithm performance in Example 2 (GLM) for SCAD penalty. The column 'avg' is the average normed differences ($\times 10^3$) between the MIST solution and the existing method's solution ; 'prop' is the proportion of MIST solutions whose objective function evaluation was less than or equal to that of the existing method's solution.

			$q = 25$						$q = 75$					
$\boldsymbol{\beta}^{(0)}$			$\mathbf{0}_p$		$\widehat{\boldsymbol{\beta}}_{ml}$		$\widehat{\boldsymbol{\beta}}_{1S,\lambda}$		$\mathbf{0}_p$		$\widehat{\boldsymbol{\beta}}_{ml}$		$\widehat{\boldsymbol{\beta}}_{1S,\lambda}$	
ρ	method		avg	prop	avg	prop	avg	prop	avg	prop	avg	prop	avg	prop
$\lambda = .001$														
0	1S		26.50	0.27	0.39	1.00	0.39	1.00	31.70	0.42	0.22	1.00	0.18	1.00
	LLA		18.55	0.68	0.15	1.00	0.13	1.00	17.31	0.76	0.22	1.00	0.11	1.00
0.5	1S		33.90	0.15	0.08	1.00	0.07	1.00	35.43	0.26	0.10	1.00	0.07	1.00
	LLA		27.65	0.64	0.01	1.00	0.00	1.00	18.45	0.82	0.10	1.00	0.00	1.00
0.75	1S		56.29	0.04	0.06	1.00	0.05	1.00	42.85	0.23	0.05	1.00	0.04	1.00
	LLA		46.48	0.71	0.05	1.00	0.00	1.00	26.05	0.82	0.04	1.00	0.00	1.00
$\lambda = .01$														
0	1S		945.60	0.11	30.65	1.00	31.42	1.00	1318.20	0.02	8.61	1.00	8.61	1.00
	LLA		416.15	0.64	5.49	0.93	1.86	0.99	406.62	0.72	0.98	1.00	0.49	1.00
0.5	1S		1082.65	0.00	23.60	1.00	22.97	1.00	1088.23	0.01	5.62	1.00	5.75	1.00
	LLA		427.10	0.72	1.33	0.99	0.03	1.00	398.05	0.74	0.56	0.99	0.16	1.00
0.75	1S		1462.74	0.00	16.81	0.98	17.37	1.00	1629.73	0.00	5.53	0.99	4.97	1.00
	LLA		548.07	0.79	1.71	0.97	0.82	1.00	578.09	0.79	1.73	0.99	0.06	1.00
$\lambda = .05$														
0	1S		1845.64	0.99	501.45	1.00	530.14	1.00	9575.27	0.82	252.36	1.00	263.41	1.00
	LLA		75.94	0.99	93.46	0.73	76.33	0.98	97.80	0.91	27.73	0.96	13.86	0.99
0.5	1S		4351.14	0.33	433.10	1.00	473.27	1.00	8323.46	0.98	171.08	1.00	181.11	1.00
	LLA		394.16	0.60	125.51	0.74	74.17	0.94	106.69	0.87	15.59	0.96	9.10	1.00
0.75	1S		5041.69	0.97	359.74	1.00	379.26	1.00	7907.54	1.00	156.65	0.99	164.34	1.00
	LLA		337.48	0.90	124.48	0.67	46.58	0.91	24.37	0.98	31.31	0.95	2.19	1.00
$\lambda = .1$														
0	1S		4095.33	1.00	818.64	1.00	815.48	1.00	8626.86	1.00	834.01	1.00	832.92	1.00
	LLA		0.00	1.00	0.04	1.00	15.14	1.00	0.00	1.00	73.78	0.89	149.55	0.98
0.5	1S		4330.64	1.00	660.87	1.00	682.83	1.00	7626.58	1.00	628.29	1.00	718.12	1.00
	LLA		4.56	1.00	32.36	0.93	34.80	0.99	0.00	1.00	115.84	0.85	121.60	0.98
0.75	1S		4536.24	1.00	626.38	1.00	693.65	1.00	7457.80	1.00	550.76	1.00	618.94	1.00
	LLA		0.00	1.00	81.21	0.87	87.10	0.99	0.00	1.00	88.95	0.86	62.41	0.98
$\lambda = .2$														
0	1S		3712.07	1.00	2888.10	0.81	3712.07	1.00	4346.96	1.00	4346.96	1.00	4346.96	1.00
	LLA		0.00	1.00	0.04	1.00	0.01	1.00	0.00	1.00	0.01	1.00	0.01	1.00
0.5	1S		3768.77	1.00	3167.21	0.98	3602.53	1.00	3781.29	1.00	3781.29	1.00	3781.29	1.00
	LLA		0.00	1.00	42.80	0.99	70.75	1.00	0.00	1.00	0.01	1.00	0.01	1.00
0.75	1S		3825.82	1.00	2542.80	0.97	3076.24	1.00	4331.74	1.00	4331.74	1.00	4331.74	1.00
	LLA		0.00	1.00	404.72	0.83	387.72	0.86	0.00	1.00	0.01	1.00	0.01	1.00
$\lambda = 1$														
0	1S		54.18	1.00	54.18	1.00	54.18	1.00	61.54	1.00	61.54	1.00	61.54	1.00
	LLA		0.00	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.02	1.00	0.00	1.00
0.5	1S		40.38	1.00	40.38	1.00	40.38	1.00	49.01	1.00	49.01	1.00	49.01	1.00
	LLA		0.00	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
0.75	1S		32.85	1.00	32.85	1.00	32.85	1.00	38.36	1.00	38.36	1.00	38.36	1.00
	LLA		0.00	1.00	0.01	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

that provided by the 1S solution. Regarding the use of LLA, which also requires a starting value specification, we again examined the cases for which LLA resulted in lower objective function evaluations. For these cases, all MIST solutions were LLA solutions, and all LLA solutions were MIST solutions with the exception of one. Hence, both methods find valid, if often different, solutions, a behavior that we again attribute to the differences in paths taken towards a solution.

4.3.3 Effectiveness of SQUAREM^2

We explored the effectiveness of SQUAREM^2, defined in Section 4.2.2, when applied to several simulated datasets taken from the previous two simulation studies. Table 4.5 indicates the relative reduction in elapsed time ('RRT') and numbers of MM updates, i.e., invocations of mapping $M(\cdot)$, required for the original and SQUAREM^2-accelerated algorithms to converge for five randomly chosen simulation datasets across the five penalty functions. The SQUAREM^2 algorithm converged without difficulty in these cases and required substantially fewer MM updates than the original algorithm; the percent reduction in time was as high as 96%. We remark here that the regularity conditions imposed in Roland and Varadhan (2005) and Varadhan and Roland (2008), particularly smoothness conditions, are not satisfied in this particular class of examples. Hence, while the simulation results are certainly very promising, the question of convergence (and its associated rate) of SQUAREM^2 in this class of problems continues to remain an interesting open problem.

4.4 Example: Identifying Genes Associated with DLBCL Survival

Diffuse large-B-cell lymphoma (DLBCL) is an aggressive type of non-Hodgkin's lymphoma and is one of the most common forms of lymphoma occurring in adults. Rosenwald et al. (2002) utilized Lymphochip DNA microarrays, specialized to include genes known to be preferentially expressed within the germinal centers of lymphoid organs, to collect and analyze gene expression data from 240 biopsy samples of DLBCL tumors. For each subject, 7399 gene expression measurements were obtained. The expression profiles along with corresponding patient information can be downloaded from their supplemental website http://llmpp.nih.gov/DLBCL/. Since the original profiles

Table 4.5: Acceleration from SQUAREM^2 applied to simplified MIST algorithm for five randomly selected simulation datasets. The reduction in elapsed time is given by 'RRT', while the number of MM updates are given for the original MIST implementation and SQUAREM^2 implementation in '# orig' and '# sqm^2', respectively.

	LAS			ALAS			EN			AEN			SCAD		
Dataset	RRT	#orig	#sqm^2	RRT	#orig	#sqm^2	RRT	#orig	#sqm^2	RRT	#orig	#sqm^2	RRT	#orig	#sqm^2
LM															
$p = 35, \sigma = 1$															
62	0.67	260	62	0.81	169	44	0.63	46	26	0.82	42	23	0.91	485	68
71	0.76	221	59	0.75	163	41	0.67	49	29	0.62	44	29	0.83	302	65
86	0.67	271	68	0.70	149	44	0.67	51	29	0.75	43	26	0.93	987	104
95	0.86	317	74	0.88	187	41	0.92	49	29	0.73	46	26	0.90	500	71
88	0.88	330	68	0.87	162	41	0.78	51	29	0.77	45	26	0.90	528	77
$p = 81, \sigma = 3$															
62	0.90	2059	242	0.89	589	92	0.65	68	35	0.75	64	29	0.88	594	101
71	0.93	1426	164	0.93	838	83	0.76	77	32	0.70	71	32	0.94	2608	215
86	0.90	1351	212	0.92	956	98	0.59	77	38	0.79	69	32	0.92	1038	110
95	0.93	1500	167	0.86	367	71	0.67	72	35	0.74	68	29	0.90	663	92
88	0.92	1547	185	0.90	716	101	0.60	70	32	0.68	66	32	0.92	1798	203
GLM															
$q = 25$															
62	0.93	4928	431	0.96	6227	272	0.89	3201	359	0.93	3316	236	0.95	22044	1442
71	0.92	4195	416	0.95	5045	239	0.90	2796	281	0.94	2843	170	0.95	16225	1052
86	0.92	4488	470	0.95	5449	242	0.92	2971	257	0.93	3044	206	0.95	20133	1193
95	0.93	4553	374	0.94	5419	341	0.92	3059	269	0.95	3096	152	0.95	15250	1064
88	0.92	5212	527	0.95	6850	371	0.91	3237	314	0.94	3393	203	0.96	26477	1367
$q = 75$															
62	0.88	4334	674	0.91	3573	377	0.85	3055	575	0.90	2435	293	0.95	88994	5687
71	0.91	3805	446	0.92	3046	281	0.85	2761	536	0.89	2194	281	0.94	82615	5588
86	0.87	3615	602	0.91	2900	329	0.87	2653	434	0.92	2110	185	0.93	42652	3686
95	0.89	3870	554	0.90	3121	380	0.90	2820	338	0.89	2264	314	0.94	40002	3095
88	0.88	4177	641	0.94	3395	251	0.87	2972	482	0.91	2415	242	0.94	77484	5885

had some missing expression measurements, we used the dataset subsequently analyzed by Li and Gui (2004) which estimated the missing values using a nearest neighbor approach. During the time of followup, 138 patient deaths were observed with median death time of 2.8 years.

Rosenwald et al. (2002) used hierarchical clustering to group the genes into four gene-expression signatures: Proliferation (PS), which includes cell-cycle control and checkpoint genes, and DNA synthesis and replication genes; Major Histocompatibility Complex ClassII (MHC), which includes genes involved in antigen presentation;

Lymph Node (LNS), which includes genes encoding for known markers of monocytes, macrophages, and natural killer cells; and Germinal Center B (GCB), which includes genes that are characteristic of germinal center B cells; see Alizadeh et al. (2000) for more information on gene signatures. Based on the gene clusters, they built a Cox proportional hazards model (Cox, 1972, 1975) to predict survival outcomes in the DLBCL patients. Subsequently, this dataset has been analyzed numerous times, typically to evaluate methods related to subgroup identification and/or survival prediction (e.g., Li and Gui, 2004; Gui and Li, 2005a,b; Li and Luan, 2005; Annest et al., 2009; Engler and Li, 2009; Tibshirani, 2009).

Here, we instead focus on the performance of two different penalties, namely SCAD and MCP, with regard to the identification of genes associated with DLBCL survival. The simulation results of Zhang (2010) suggest that MCP has superior selective accuracy over the SCAD penalty, at least for the case of a linear model. There, selection accuracy was measured as the proportion of simulation replications with correct classification of both the zero and non-zero coefficients, with MCP outperforming SCAD in all simulation settings. To illustrate the utility and flexibility of the MIST algorithm, we reanalyzed the DLBCL data, fitting a penalized Cox regression model respectively using SCAD and MCP penalty functions, and running these procedures in combination with the Iterative Sure Independence Screening procedure (ISIS, Fan et al., 2009b) in order to ensure that the dimension of the parameter space was maintained at a manageable level. For SCAD, we considered both the 1S and LLA estimators. The existing optimization functions provided in the *SIS* package for the ISIS procedure were used for the 1S estimator, whereas relevant modifications to the ISIS code were made in order to accommodate the fully iterative LLA and MCP estimators. Optimization at each step of the ISIS algorithm in the case of the MCP penalty utilized the MIST algorithm, as we are aware of no other algorithm capable of fitting the Cox regression model subject

to MCP penalization. The default settings in the *SIS* package were used to determine the maximum number of predictors ($[\frac{n}{4 \log n}] = 10$) and to define the additional ISIS parameters (e.g., use of the unpenalized MLE as a starting value, ranking method, tuning parameter selection) for all three analyses (1S-SCAD, LLA-SCAD, MIST-MCP). The choice $a = 3.7$ was used for all analyses; hence, only the selection of λ required tuning.

Table 4.6 displays the 11 genes identified by at least one of the three analyses. The x's in a given column indicate the genes with non-zero coefficients resulting from the corresponding penalization. The final column provides references for genes previously linked to DLBCL in the literature. Genes belonging to the original Rosenwald et al. (2002) gene expression signatures are indicated with parenthetical initials. Note that the references provided are not meant to be an exhaustive list, but instead to demonstrate the relevance of certain genes and/or their altered expression levels in DLBCL survival.

Interestingly, the LLA-SCAD and MIST-MCP penalizations selected the same subset of genes, with a nearly a complete overlap with those selected from the 1S-SCAD penalization. The number of genes selected in each case is 10, the maximum specified by ISIS; 9 of these were shared across the three penalizations. According to NCBI Entrez Gene search (http://www.ncbi.nlm.nih.gov/), many of these genes are biologically relevant. For example, CDK7 codes for a protein that regulates cell cycle progression and is represented in the Proliferation Signature, although reported under a different Lymphochip ID as this gene was spotted multiple times on the array. Also members of the Proliferation Signature are SEPT1, coding for a protein involved in cytokinesis, and BUB3, coding for a mitotic checkpoint protein. DNTTIP2 regulates transcriptional activity of DNTT, a gene for a protein expressed in a restricted population of normal and malignant pre-B and pre-T lymphocytes during early differentiation. HLA-DRA, a member of the MHC Signature, plays a central role in the immune sys-

tem and is expressed in antigen presenting cells, such as B lymphocytes, dendritic cells, macrophages. From the GCB Signature, the ESTs weakly similar to thyroxine-binding globulin precursor is highly cited. Additionally, RFTN1 plays a pivotal role in regulating B-cell antigen receptor-mediated signaling (Saeki et al., 2003).

A description of AI568329 was not provided in the original dataset, thus its function is unknown. Similarly, although cited at least twice, a description for AA830781 was also not provided in the original dataset. However, both of these may be related to lymphoma or risk of death from lymphoma, as it is possible that these genes (and potentially others) were selected because of coexpression or correlation with other relevant genes.

Interestingly the two genes not identified across all penalizations were both cited in Martinez-Climent et al. (2003). They found altered expression of TSC22D3 and ITGAL (both involved in various immune phenomena) in one case who initially presented with follicle center lymphoma and subsequently transformed to DLBCL.

The results of this analysis demonstrate equivalence in selection performance between MCP and LLA-SCAD for the case of Cox proportional hazards model. Increasing the maximum number of predictors to 21 again resulted in equivalent selection performance between MCP and LLA-SCAD, with 21 predictors ultimately selected (results not shown). The 1S estimator also resulted in the selection of 21 predictors, but with increased dissimilarity between MCP/LLA-SCAD and 1S: only 13 of the 21 genes were selected by all three methods. It should be noted that Zhang (2010) did not use any form iterative variable selection (e.g., ISIS) in his comparisons between SCAD and MCP for the case of the linear model; in addition, Zhang (2010) fixed values for both penalty parameters in his simulations and also did not use $a = 3.7$. Thus, the ISIS procedure, the tuning parameter selection process, and/or the choice of $a = 3.7$ (as suggested in Fan et al. (2009b)) could all play a role in the results reported here.

Table 4.6: Genes associated with DLBCL survival with SCAD (one-step=1S and LLA) and MCP penalizations, sorted by the gene order in the original data set. ID refers to the unique Lymphochip identification number. The x's in a given column indicate the genes identified by the corresponding penalization.

ID	Name (Symbol)	SCAD 1S	SCAD LLA	MCP	References
27774	cyclin-dependent kinase 7 (CDK7)	x	x	x	Rosenwald et al. (2002) (PS) Ma and Huang (2007) Binder and Schumacher (2008, 2009)
31242	acidic 82 kDa protein mRNA (DNTTIP2)	x	x	x	Binder and Schumacher (2008, 2009)
31981	septin 1 (SEPT1)	x	x	x	Rosenwald et al. (2002) (PS) Li and Luan (2005) Sinisi et al. (2006), Sha et al. (2006) Zhang and Zhang (2007) Annest et al. (2009) Binder and Schumacher (2008, 2009)
29652	BUB3 budding uninhibited by benzimidazoles 3 (BUB3)	x	x	x	Rosenwald et al. (2002) (PS)
27731	major histocompatibility complex, class II, DR alpha (HLA-DRA)	x	x	x	Rosenwald et al. (2002) (MHC) Li and Luan (2005) Gui and Li (2005a,b) Sohn et al. (2009) Binder and Schumacher (2009)
24376	ESTs, Weakly similar to A47224 thyroxine-binding globulin precursor	x	x	x	Rosenwald et al. (2002) (GCB) Ando et al. (2003) Gui and Li (2005a,b) Li and Luan (2005) Annest et al. (2009) Sohn et al. (2009) Binder and Schumacher (2008, 2009)
22162	delta sleep inducing peptide, immunoreactor (TSC22D3)		x	x	Martinez-Climent et al. (2003)
23862	(AI568329) ESTs	x	x	x	
24271	integrin, alpha L (ITGAL)	x			Martinez-Climent et al. (2003)
33358	(AA830781)	x	x	x	Li and Luan (2005) Binder and Schumacher (2009)
32679	KIAA0084 protein (RFTN1)	x	x	x	Gui and Li (2005b), Sha et al. (2006) Zhang and Zhang (2007) Annest et al. (2009) Binder and Schumacher (2008, 2009)

4.5 Proofs of Theorems

Proof of Theorem 4.1.1: The assumptions stated in the theorem immediately yield that $\xi(\boldsymbol{\beta})$ is locally Lipschitz continuous and coercive for each bounded $\lambda > 0$, hence (i) is satisfied.

To show (ii), we first write

$$\begin{aligned} q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda}) &= \sum_{j=1}^{p} \left[\tilde{q}(|\beta_j|, |\alpha_j|; \boldsymbol{\lambda}_j) - \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j) \right] \\ &= \sum_{j=1}^{p} \left[\tilde{p}(|\alpha_j|; \boldsymbol{\lambda}_j) + \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)(|\beta_j| - |\alpha_j|) - \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j) \right]. \end{aligned} \tag{4.19}$$

This difference is obviously equal to zero whenever $\boldsymbol{\beta} = \boldsymbol{\alpha}$. For $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$, we shall separately consider the case where $\tilde{p}(r; \boldsymbol{\lambda}_j)$ is linear versus nonlinear.

First, suppose that $\tilde{p}(r; \boldsymbol{\theta}) = a_1 + a_2 r$, where $a_1 \geq 0$ and $a_2 > 0$ and each may depend on $\boldsymbol{\theta}$. It then follows immediately that

$$\begin{aligned} \tilde{p}(|\alpha_j|; \boldsymbol{\lambda}_j) &+ \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)(|\beta_j| - |\alpha_j|) - \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j) \\ &= (a_1 + a_2|\alpha_j|) + a_2(|\beta_j| - |\alpha_j|) - (a_1 + a_2|\beta_j|) = 0. \end{aligned}$$

Thus, the claimed equality between (4.1) and (4.2) holds in this case.

Now, suppose that $\tilde{p}(r; \boldsymbol{\theta})$ is nonlinear in r. Under (P1), we claim that (4.2) strictly majorizes $p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ provided the derivative of the penalty $\tilde{p}'(\cdot, \boldsymbol{\lambda}_j)$ is strictly positive. To see this, observe that concavity (e.g., see (4.4)) implies the inequality

$$\tilde{q}(r, s; \boldsymbol{\theta}) - \tilde{p}(r; \boldsymbol{\theta}) = -1 \left[\tilde{p}(r; \boldsymbol{\theta}) - \tilde{p}(s; \boldsymbol{\theta}) - \tilde{p}'(s; \boldsymbol{\theta})(r - s) \right] \geq 0,$$

with equality holding if and only if $r = s$ and $p'(s; \boldsymbol{\theta}) > 0$. For penalties such that their derivatives are nonnegative, i.e., $p'(s; \boldsymbol{\theta}) \geq 0$, we obtain the same inequality as above, with equality additionally holding for r and s sufficiently large. Therefore,

$$q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \sum_{j=1}^{p} \left[\tilde{q}(|\beta_j|, |\alpha_j|; \boldsymbol{\lambda}_j) - \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j) \right] \geq 0,$$

and (ii) is established.

In order to establish the majorization property specified in (iii), we begin by noting that our assumptions on $g(\boldsymbol{\beta})$, $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$, and $\tilde{p}(\cdot; \boldsymbol{\theta})$ imply that $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and

$\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ are both continuous in $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Our assumptions further imply that $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) \geq 0$; if at least one of $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ or $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is strictly positive for $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$, then $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 0$ for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ and $\psi(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$. Therefore, the objective function $\xi(\boldsymbol{\beta})$ is strictly majorized by $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \equiv \xi(\boldsymbol{\beta}) + \psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

In order to establish the convergence of the corresponding MM algorithm in (iii), it suffices to prove that the assumptions of the theorem and consequent assertions established thus far are sufficient to ensure that Conditions R1-R5 of Section 3.1 are met, in which case Theorem 3.1.3 applies directly. The result (i), combined with the assumption that the stationary points are all isolated, immediately establishes Condition R1; as proved above, conditions R2 and R3 also hold. If $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is continuous in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and locally Lipschitz continuous in $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$, then (i) implies that R4 also holds. By assumption, $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is continuous in $\boldsymbol{\alpha}$ and continuously differentiable in $\boldsymbol{\beta}$, hence locally Lipschitz in $\boldsymbol{\beta}$. Continuity of $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ in both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is also immediate. Hence, R4 holds provided that $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is locally Lipschitz continuous in $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$. To see that this is the case, we note that (4.19) is a linear combination of functions in β_j of the form $\tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j| - \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j)$, where $|\cdot|$ and $-\tilde{p}(\cdot; \boldsymbol{\lambda})$ are both convex, hence locally Lipschitz. Since both the sum and composition of two locally Lipschitz functions are locally Lipschitz, the result now follows. Finally, R5 is ensured by R1-R4 and the condition in (iii) that $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is uniquely minimized in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha}$. □

Remark 4.5.1. *Under the conditions of this theorem, one may appeal to the to the theory of the Clarke subdifferential summarized in Section 2.3 and prove directly that the stationary points of $\xi(\boldsymbol{\beta})$ (i.e., defined in the sense of Clarke (1990)) coincide with the fixed points of $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$. In particular, since $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is locally Lipschitz continuous for $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$ for each bounded $\boldsymbol{\alpha}$, the relation $\boldsymbol{\beta}^* = M(\boldsymbol{\beta}^*)$ is equivalent to*

$$0 \in \partial \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*},$$

where the right-hand side denotes the Clarke subdifferential of $\xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ *taken with respect to* $\boldsymbol{\beta}$*, evaluated at* $\boldsymbol{\beta} = \boldsymbol{\beta}^*$*. By result C7 in Section 2.3,*

$$\partial \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \subset \partial \xi(\boldsymbol{\beta}^*) \oplus \partial \psi(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}.$$

It follows that $\boldsymbol{\beta}^*$ *is a stationary point of* $\xi(\boldsymbol{\beta})$ *provided only that* $\partial \psi(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \{0\}$*, since in this case we have*

$$\partial \xi^{SUR}(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \equiv \partial \xi(\boldsymbol{\beta}^*)$$

and hence that $0 \in \partial \xi(\boldsymbol{\beta}^*)$*.*

To establish that $\partial \psi(\boldsymbol{\beta}, \boldsymbol{\beta}^*)|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = \{0\}$*, we recall that* $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *is locally Lipschitz continuous in* $\boldsymbol{\beta}$ *for* $\boldsymbol{\beta}$ *near* $\boldsymbol{\alpha}$ *and additionally satisfies* $\psi(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$ *and* $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 0$ *for* $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ *under the conditions specified in (iii). The corresponding full set of assumptions on* $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *clearly imply that the derivative of* $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *must exist in* $\boldsymbol{\beta}$ *and equal zero at* $\boldsymbol{\beta} = \boldsymbol{\alpha}$*. Using result C7 from Section 2.3, it follows that the Clarke subdifferential of* $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *is* $\{\mathbf{0}\}$ *at* $\boldsymbol{\beta} = \boldsymbol{\alpha}$ *for every* $\boldsymbol{\alpha}$ *if the subdifferential of* $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$*, considered as a function of* $\boldsymbol{\beta}$ *for each* $\boldsymbol{\alpha}$*, is also* $\{\mathbf{0}\}$ *at* $\boldsymbol{\beta} = \boldsymbol{\alpha}$*. This is trivially satisfied if* $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ *is identically zero. Hence, we must only consider what happens in the case where* $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ *is zero at* $\boldsymbol{\beta} = \boldsymbol{\alpha}$ *and nonnegative otherwise. In this case, the result that the subdifferential of* $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$*, is* $\{\mathbf{0}\}$ *at* $\boldsymbol{\beta} = \boldsymbol{\alpha}$ *follows directly from* (4.2) *and the following proposition, whose proof is also provided below.*

Proposition 4.5.2. *For every finite* s *and bounded* $\boldsymbol{\theta}$ *such that the first element is* $\lambda > 0$ *and the remainder of the elements the additional parameters defining the penalty, we have*

$$\partial \left[\tilde{q}(|r|, |s|; \boldsymbol{\theta}) - \tilde{p}(|r|; \boldsymbol{\theta}) \right] = \varphi(|r|, |s|; \boldsymbol{\theta}) \times \partial |r|,$$

where the operation on the right hand side denotes the multiplication of every element

in the subdifferential $\partial|r|$ by the scalar $\varphi(|r|, |s|; \boldsymbol{\theta})$ and, for $u, v \geq 0$,

$$\varphi(u, v; \boldsymbol{\theta}) = \tilde{p}'(v; \boldsymbol{\theta}) - \tilde{p}'(u; \boldsymbol{\theta}).$$

In particular, $\partial \left[\tilde{q}(|r|, |s|; \boldsymbol{\theta}) - \tilde{p}(|r|; \boldsymbol{\theta}) \right] = \{0\}$ at $r = s$.

Proof: *Define $f_1(u) = \tilde{q}(u, |s|; \boldsymbol{\theta}) - \tilde{p}(u; \boldsymbol{\theta})$ for $u \geq 0$ and $f_2(v) = |v|$; then,*

$$\tilde{q}(|r|, |s|; \boldsymbol{\theta}) - \tilde{p}(|r|; \boldsymbol{\theta}) = f_1(f_2(r)).$$

Result C8 in Section 2.3 now implies the proposition if $f_1(u)$ is continuously differentiable and $f_1'(u) = \tilde{p}'(|s|; \boldsymbol{\theta}) - \tilde{p}'(u; \boldsymbol{\theta})$ for $u \geq 0$. However, this is immediate; since

$$f_1(u) = \tilde{p}(|s|; \boldsymbol{\theta}) - \tilde{p}(u; \boldsymbol{\theta}) + \tilde{p}'(|s|; \boldsymbol{\theta})(u - |s|),$$

the fact that $\tilde{p}(u; \boldsymbol{\theta})$ is continuously differentiable for $u \geq 0$ implies

$$f_1'(u) = -\tilde{p}'(u; \boldsymbol{\theta}) + \tilde{p}'(|s|; \boldsymbol{\theta}),$$

with the derivatives in the last expression taken to be right derivatives for $u = s = 0$. □

Proof of Theorem 4.2.2: Under the stated conditions and for any bounded $\boldsymbol{\alpha}$, $m(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2$ is strictly convex with a Lipschitz continuous derivative of order $L^{-1} > 0$; in addition, $\sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|$ is also convex in $\boldsymbol{\beta}$. Hence, for each bounded $\boldsymbol{\alpha}$ there exists a unique solution $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\boldsymbol{\alpha})$ when minimizing (4.8).

In the notation of Combettes and Wajs (2005), we may identify the Hilbert space $\mathcal{H}$ with $\mathbb{R}^p$, $f_2(\boldsymbol{\beta})$ with $m(\boldsymbol{\beta})$ and $f_1(\boldsymbol{\beta})$ with $\sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|$. The assumptions of the theorem ensure that the regularity conditions of Proposition 3.1 and Theorem 3.4 of Combettes and Wajs (2005) are met. In particular, because $m(\boldsymbol{\beta})$ is coercive and strictly convex, Proposition 3.1 guarantees the existence of a unique solution to

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} f_1(\boldsymbol{\beta}) + f_2(\boldsymbol{\beta})$$

as well as provides the relevant fixed point mapping; Theorem 3.4 establishes weak convergence of the corresponding iterative scheme to this unique solution. Since weak convergence is equivalent to strong convergence in a finite dimensional Hilbert space, such results imply componentwise convergence of the iteration sequence to $\boldsymbol{\beta}^*$.

Both Proposition 3.1 and Theorem 3.4 of Combettes and Wajs (2005) rely on the gradient of $f_2(\boldsymbol{\beta})$ and the so-called "proximity operator" of $f_1(\boldsymbol{\beta})$. Example 2.20 in Combettes and Wajs (2005) shows that the proximity operator for $f_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|$ is exactly $S(\cdot; \boldsymbol{\tau})$. The algorithm summarized in the statement of the theorem is therefore observed to be a specific instance of that described in the Theorem 3.4 with (in their notation) $a_n = b_n = 0$ and $\lambda_n = 1$ for every n.

Hale et al. (2008, Theorem 4.5) undertake a detailed study of the proposed algorithm for the special case of a convex, differentiable $f_2(\boldsymbol{\beta})$ and $f_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$. In this case, they prove that the algorithm converges in a finite number of iterations.

A minor extension of their arguments may be used to establish the same result for $f_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|$, provided that $\tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j) \in [0, \infty)$ for each j. □

Proof of Theorem 4.2.4: To establish (1.), note that the choice of $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ in (4.11) with appropriate ϖ guarantees majorization of $-\ell(\widetilde{\boldsymbol{\beta}})$ provided $\nabla^2(-\ell(\widetilde{\boldsymbol{\beta}}))$ can be bounded (e.g., Lange, 2004, Ch 6). Penalties of form (4.1) satisfying assumption (P1) can be linearly majorized so that (4.12) majorizes $\xi_{glm}(\tilde{\boldsymbol{\beta}})$. For (2.), $-\ell(\widetilde{\boldsymbol{\beta}})$ is indeed strictly convex and coercive, with $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) \geq 0$ continuous in both $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\alpha}}$ and continuously differentiable in $\widetilde{\boldsymbol{\beta}}$ for each $\widetilde{\boldsymbol{\alpha}}$, with $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = 0$ when $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\alpha}}$. As for (3.), let $\zeta = 2\varpi^{-1}$. Note that the surrogate $\xi^{SUR}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ is differentiable in β_j only if $\beta_j \neq 0$. Assuming $\beta_j \neq 0$, $j \neq 0$ and excluding irrelevant constants,

$$\frac{\partial \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}; \tilde{\boldsymbol{\alpha}})}{\partial \beta_j} = -\left[\nabla \ell(\tilde{\boldsymbol{\alpha}})\right]_j + \zeta\beta_j - \zeta\alpha_j + \tau_j \text{sign}(\beta_j) + 2\lambda\varepsilon\beta_j. \tag{4.20}$$

Setting (4.20) equal to zero implies

$$\beta_j = \begin{cases} \frac{1}{\zeta+2\lambda\varepsilon}\left(\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_j + \zeta\alpha_j - \tau_j\right) & \beta_j > 0 \\ \frac{1}{\zeta+2\lambda\varepsilon}\left(\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_j + \zeta\alpha_j + \tau_j\right) & \beta_j < 0 \end{cases}.$$

For sign consistency, we impose that $\frac{1}{\zeta+2\lambda\varepsilon}\left(\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_j + \zeta\alpha_j\right) > \tau_j$ when $\beta_j > 0$ and $\frac{1}{\zeta+2\lambda\varepsilon}\left(\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_j + \zeta\alpha_j\right) < -\tau_j$ when $\beta_j < 0$. When $\left|\frac{1}{\zeta+2\lambda\varepsilon}\left(\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_j + \zeta\alpha_j\right)\right| \leq \tau_j$, we set $\beta_j = 0$. In summary,

$$\beta_j^* = \frac{1}{\zeta+2\lambda\varepsilon} s\left(\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_j + \zeta\alpha_j, \tau_j\right),$$

from which the first part of (4.13) directly follows for $j \in \{1,\ldots,p\}$. We do not penalize the intercept, thus

$$\frac{\partial \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}};\tilde{\boldsymbol{\alpha}})}{\partial\beta_0} = -\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_0 + \zeta\beta_0 - \zeta\alpha_0$$

so that $\beta_0^* = (\left[\nabla\ell(\tilde{\boldsymbol{\alpha}})\right]_0 + \zeta\alpha_0)/\zeta$.

Furthermore, take $\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\kappa}}$ for any $\tilde{\boldsymbol{\beta}} \in \mathcal{R}^{p+1}$ and $\tilde{\boldsymbol{\kappa}} = (\kappa_0, \boldsymbol{\kappa}^T)^T \in \mathcal{R}^{p+1}$ is arbitrary. Then, following arguments similar to those in Daubechies et al. (2004, Prop. 2.1),

$$\begin{aligned} \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}+\tilde{\boldsymbol{\kappa}},\tilde{\boldsymbol{\alpha}}) &= -\ell(\tilde{\boldsymbol{\alpha}}) - \nabla\ell(\tilde{\boldsymbol{\alpha}})'(\tilde{\boldsymbol{\beta}}+\tilde{\boldsymbol{\kappa}}-\tilde{\boldsymbol{\alpha}}) + \frac{\zeta}{2}(\tilde{\boldsymbol{\beta}}+\tilde{\boldsymbol{\kappa}}-\tilde{\boldsymbol{\alpha}})'(\tilde{\boldsymbol{\beta}}+\tilde{\boldsymbol{\kappa}}-\tilde{\boldsymbol{\alpha}}) \\ &\quad + \sum_{j=1}^{p}(\tau_j|\beta_j+\kappa_j| + \gamma_j + \lambda\varepsilon(\beta_j+\kappa_j)^2) \\ &= \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}) + (\frac{\zeta}{2}+\lambda\varepsilon)\boldsymbol{\kappa}'\boldsymbol{\kappa} + \frac{\zeta}{2}\kappa_0^2 \\ &\quad + \kappa_0(\zeta\beta_0 - \zeta\alpha_0 - [\nabla\ell(\tilde{\boldsymbol{\alpha}})]_0) + \sum_{j=1}^{p}\left[\tau_j(|\beta_j+\kappa_j| - |\beta_j|)\right. \\ &\quad \left. + \kappa_j((\zeta+2\lambda\varepsilon)\beta_j - \zeta\alpha_j - [\nabla\ell(\tilde{\boldsymbol{\alpha}})]_j)\right]. \end{aligned} \tag{4.21}$$

Consider $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}^* \equiv [\beta_0^*, \boldsymbol{\beta}^{*T}]^T$ where $\tilde{\boldsymbol{\beta}}^*$ defined in (4.13), and define sets $\mathcal{J} = \{1,2,\ldots,p\}$, $\mathcal{J}_0 = \{j \in \mathcal{J} : \beta_j^* = 0\}$ and $\mathcal{J}_1 = \mathcal{J}\backslash\mathcal{J}_0$. Noting that β_j^* satisfies $(\zeta+2\lambda\varepsilon)\beta_j^* - \zeta\alpha_j - [\nabla\ell(\boldsymbol{\alpha})]_j = -\tau_j\text{sign}(\beta_j^*)$ for $j \in \mathcal{J}_1$, and noting that

$\zeta\beta_0^* - \zeta\alpha_0 - [\nabla\ell(\tilde{\boldsymbol{\alpha}})]_0 = 0$, we have

$$\begin{aligned}\xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}^* + \tilde{\boldsymbol{\kappa}}, \tilde{\boldsymbol{\alpha}}) &- \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\alpha}}) \\ &= (\frac{\zeta}{2} + \lambda\varepsilon)\boldsymbol{\kappa}'\boldsymbol{\kappa} + \frac{\zeta}{2}\kappa_0^2 + \sum_{j\in\mathcal{J}_0}\left[\tau_j|\kappa_j| - \kappa_j(\zeta\alpha_j + [\nabla\ell(\boldsymbol{\alpha})]_j)\right] \\ &\quad + \sum_{j\in\mathcal{J}_1}\left[\tau_j(|\beta_j^* + \kappa_j| - |\beta_j^*|) - \kappa_j\tau_j\text{sign}(\beta_j^*)\right].\end{aligned}$$

For $j \in \mathcal{J}_0$, $|\zeta\alpha_j + [\nabla\ell(\widetilde{\boldsymbol{\alpha}})]_j| \leq \tau_j$, so that $\tau_j|\kappa_j| - \kappa_j(\zeta\alpha_j + [\nabla\ell(\widetilde{\boldsymbol{\alpha}})]_j) \geq 0$. For $j \in \mathcal{J}_1$, there are two cases, corresponding to the sign of β_j^*. First consider $\beta_j^* > 0$, then

$$\tau_j(|\beta_j^* + \kappa_j| - |\beta_j^*|) - \kappa_j\tau_j\text{sign}(\beta_j^*) \;=\; \tau_j(|\beta_j^* + \kappa_j| - (\beta_j^* + \kappa_j)) \geq 0.$$

If $\beta_j^* < 0$, then

$$\tau_j(|\beta_j^* + \kappa_j| - |\beta_j^*|) - \kappa_j\tau_j\text{sign}(\beta_j^*) \;=\; \tau_j(|\beta_j^* + \kappa_j| + (\beta_j^* + \kappa_j)) \geq 0.$$

Thus, $\xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}^* + \tilde{\boldsymbol{\kappa}}, \tilde{\boldsymbol{\alpha}}) - \xi_{glm}^{SUR}(\tilde{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\alpha}}) \geq (\frac{\zeta}{2} + \lambda\varepsilon)\boldsymbol{\kappa}'\boldsymbol{\kappa} + \frac{\zeta}{2}\kappa_0^2 \geq \frac{\zeta}{2}\tilde{\boldsymbol{\kappa}}'\tilde{\boldsymbol{\kappa}}$, since $\lambda\varepsilon \geq 0$, hence guaranteeing a unique minimum, and proving the proposition. □

CHAPTER 5

MIST AND FINITE MIXTURE REGRESSION MODELS

In regression modeling, the goal is to relate a response variable y to a set of covariates $\mathbf{x} = (x_1, \ldots, x_p)$. The usual approach requires estimating a single set of regression coefficients, shared by all of the observed samples $(y_1, \mathbf{x}_1), \ldots, (y_N, \mathbf{x}_N)$. It is often the case, especially with a large number of covariates, that the N observed samples are not adequately modeled using the same set of regression coefficients; that is, a set or subset of coefficients may be different for different subgroups of observations. Additionally, it may be possible for some coefficients in some (or all) subgroups to be zero.

In this chapter, finite mixture regression (FMR) model fitting is explored when the number of components is potentially unknown and the regression coefficients within each component are allowed to be differentially sparse. As such, a Majorization-Minimization (MM) procedure is proposed, which encompasses an expectation majorization step as in the EM algorithm (treated as an Expectation *Minimization* algorithm in this context) to estimate the mixture parameters and MIST to estimate the penalized regression coefficients. Motivated by the work of Städler et al. (2010), Block Coordinate Descent (BCD) is used to estimate the parameters within the MM algorithm. Importantly, the incorporation of MIST allows for more general, penalized forms of finite mixture regression models than considered in Städler et al. (2010).

The chapter begins with a review of unpenalized and penalized finite mixture regression models, followed by the proposed MIST-MIX optimization algorithm and details for the linear mixture model setting with unknown common variance. The convergence of the algorithm is then discussed, with proofs relegated to the end of the chapter. Simulation results and analysis of the well-known ozone meteorological dataset (Breiman and Friedman, 1985) are also provided.

5.1 Unpenalized Finite Mixture Regression Model

Consider the traditional (unpenalized) finite mixture regression model

$$f(\mathbf{y};\boldsymbol{\phi}) = \prod_{i=1}^{N}\sum_{r=1}^{K}\pi_r f_r(y_i;\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_r), \tag{5.1}$$

where $\mathbf{y} = (y_1,\ldots,y_N)^T$, $\boldsymbol{\phi} = \{(\pi_r,\widetilde{\boldsymbol{\beta}}_r) : r = 1,\ldots,K\}$, $0 \le \pi_r \le 1$, $\sum_{r=1}^{K}\pi_r = 1$, $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ is a convex subset of $\mathbb{R}^P$ with P denoting the length of $\boldsymbol{\phi}$, $f_r(y_i;\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_r)$ is a density dependent on $p+1$ regression coefficients $\widetilde{\boldsymbol{\beta}}_r = (\beta_{r0},\beta_{r1},\ldots,\beta_{rp})^T = (\beta_{r0},\boldsymbol{\beta}_r)^T$ of covariates $\tilde{\mathbf{x}}_i$ comprising the rows of the $N \times (p+1)$ design matrix $\widetilde{\mathbf{X}} = [\mathbf{1}_N,\mathbf{X}]$, and K is the *maximum* possible number of mixture components considered; i.e., some components r may have $\pi_r = 0$. Further suppose the same number of features (p) are being considered for each mixture component, with intercepts allowed to differ across components. The most common density choice for f_r is the normal density, in which case one or more scale parameter should also be included in $\boldsymbol{\phi}$.

Estimation of the parameters $\boldsymbol{\phi}$ is typically achieved using the EM algorithm, with component membership serving as the "missing data." In the general paradigm, one typically wishes to maximize the loglikelihood $\log f(\mathbf{y};\boldsymbol{\phi})$ in the parameters $\boldsymbol{\phi}$, where $f(\mathbf{y};\boldsymbol{\phi})$ is given in (5.1). The complete-data loglikelihood that corresponds to (5.1) is obtained by assuming group membership is known, and is given by

$$\ell_C(\boldsymbol{\phi}) \;=\; \sum_{i=1}^{N}\sum_{r=1}^{K} z_{ir}\{\log f_r(y_i;\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_r) + \log(\pi_r)\} \tag{5.2}$$

where $z_{ir} = 1$ when the i^{th} observation comes from the r^{th} mixture component and 0 otherwise. At iteration k, the E-step involves taking the expectation of (5.2) so that z_{ir} is replaced with its expected value

$$\delta_{ir}^{(k)} = \frac{\pi_r^{(k)} f(y_i;\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_r^{(k)})}{\sum_{m=1}^{K}\pi_m^{(k)} f(y_i;\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_m^{(k)})},\; r = 1,\ldots,K,\; i = 1,\ldots,N, \tag{5.3}$$

while the M-step maximizes the resulting surrogate function.

5.2 Penalized Finite Mixture Regression Model

To induce sparsity in the regression coefficients, one may include a nonsmooth penalty function like those discussed in Chapter 4 to the (scaled) negative loglikelihood. Recently, Khalili and Chen (2007) proposed an EM-based algorithm for FMR in the spirit of Hunter and Li (2005) using a perturbed (differentiable) penalty function, while Städler et al. (2010) proposed a generalized EM algorithm for L_1-penalized linear (normal) finite mixture regression models. The MM algorithm, and in particular, MIST, can be used to minimize more general, penalized forms of finite mixture regression models while capitalizing on the sparsity inherent to singular penalties at the origin. Consider the objective function

$$\xi(\boldsymbol{\phi}) = \underbrace{-\frac{1}{N}\sum_{i=1}^{N}\log\sum_{r=1}^{K}\pi_r f_r(y_i;\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_r)}_{g_N(\boldsymbol{\phi})} + \sum_{r=1}^{K}\pi_r^{\upsilon}\{p(\boldsymbol{\beta}_r;\boldsymbol{\lambda}_r) + \lambda\varepsilon\|\boldsymbol{\beta}_r\|^2\}, \qquad (5.4)$$

where $g_N(\boldsymbol{\phi}) = -N^{-1}\ell(\boldsymbol{\phi})$ is the scaled negative loglikelihood, and the penalty $p(\boldsymbol{\beta}_r;\boldsymbol{\lambda}_r)$ is the sum of separable components, $\tilde{p}(|\beta_{rj}|;\boldsymbol{\lambda}_{rj})$, which satisfy condition (P1) from Chapter 4 for each $r = 1,\ldots,K$. The final term in (5.4) allows for elastic-net type penalties for $\varepsilon > 0$. As in Khalili and Chen (2007) and Städler et al. (2010), we incorporate a multiplier π_r^{υ} for the penalty, which in effect, scales the penalty differentially depending on the size of the group r. Since the density for group r is also weighted by π_r, the inclusion of π_r in the penalty (with $\upsilon = 1$) may be thought of as an equalization measure that puts the penalty on the same scale relative to the density. Whereas Khalili and Chen (2007) only consider $\upsilon = 1$, Städler et al. (2010) consider $\upsilon \in \{0, 1/2, 1\}$ in simulations, but only $\upsilon = 0$ in their proofs. Based on their simulation results, Städler et al. (2010) comment that all three values of υ perform similarly for balanced cases (equal number of observations within groups), but $\upsilon = 1$ performed more favorably when the groups were unbalanced. For this reason, we consider $\upsilon \in \{0, 1\}$. For com-

patibility with MIST, attention is restricted to choices $f_r(\cdot) = f(\cdot)$ corresponding to canonically parameterized generalized linear models with bounded hessians. Thus, the probability distribution of **y** follows a generalized linear model with a canonical link and linear predictor $\widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}$. While a finite mixture of linear regressions with different unknown scale parameters for each component is possible within this specification, the case of common unknown variance is considered separately as an example in Section 5.2.2.

The proposed surrogate majorizing function of (5.4) results from majorizing the different portions of the objection function:

1. EM majorization. Recall from Section 2.2 that $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) = -\ell(\boldsymbol{\phi}) - H(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$, where $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) + H(\boldsymbol{\phi}^{(k)}, \boldsymbol{\phi}^{(k)}) = -\ell(\boldsymbol{\phi}) + D(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$. Here, $Q(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) = -E_{\boldsymbol{\phi}^{(k)}}(\ell_C(\phi)|\mathbf{y})$, where $\ell_C(\phi)$ is given in (5.2). This implies

$$\begin{aligned} \xi(\boldsymbol{\phi}) &\leq g_N(\boldsymbol{\phi}) + D_N(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) + \sum_{r=1}^{K} \pi_r^{\upsilon} \left\{ p(\boldsymbol{\beta}_r; \boldsymbol{\lambda}_r) + \lambda\varepsilon\|\boldsymbol{\beta}_r\|^2 \right\} \\ &= \xi(\boldsymbol{\phi}) + D_N(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}), \end{aligned}$$

   where $D_N(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) = N^{-1} D(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$.

2. 'Separation' majorization in $\widetilde{\boldsymbol{\beta}}$. Define $J(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = \sum_{r=1}^{K} h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r)$, where $h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) \geq 0$ is a real-valued, continuous function of $\widetilde{\boldsymbol{\beta}}_r$ and $\widetilde{\boldsymbol{\alpha}}_r$ that is continuously differentiable in $\widetilde{\boldsymbol{\beta}}_r$ for each $\widetilde{\boldsymbol{\alpha}}_r$ and satisfies $h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) = 0$ when $\widetilde{\boldsymbol{\beta}}_r = \widetilde{\boldsymbol{\alpha}}_r$ for $r = 1, \ldots, K$. Thus, $J(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) \geq 0$ and satisfies $J(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = 0$ when $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\alpha}}$, so that for $\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\beta}}^{(k)}$

$$\begin{aligned} \xi(\boldsymbol{\phi}) &\leq g_N(\boldsymbol{\phi}) + D_N(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) + J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}^{(k)}) \\ &\quad + \sum_{r=1}^{K} \pi_r^{\upsilon} \left\{ p(\boldsymbol{\beta}_r; \boldsymbol{\lambda}_r) + \lambda\varepsilon\|\boldsymbol{\beta}_r\|^2 \right\} \\ &= \xi(\boldsymbol{\phi}) + D_N(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) + J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}^{(k)}), \end{aligned}$$

   where $J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}^{(k)}) = N^{-1} J(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}}^{(k)})$.

3. Penalty majorization. Majorize $p(\boldsymbol{\beta}_r;\boldsymbol{\lambda}_r)$ linearly through $q(\boldsymbol{\beta}_r,\boldsymbol{\alpha}_r;\boldsymbol{\lambda}_r)$ for any arbitrary (bounded) $\boldsymbol{\alpha}_r$ provided the penalty function $\tilde{p}$ satisfies (P1). That is,

$$q(\boldsymbol{\beta}_r,\boldsymbol{\alpha}_r;\boldsymbol{\lambda}_r)=\sum_{j=1}^{p}\tilde{q}(|\beta_{rj}|,|\alpha_{rj}|;\boldsymbol{\lambda}_{rj}), \tag{5.5}$$

where $\tilde{q}(t,s;\boldsymbol{\theta})=\tilde{p}(s;\boldsymbol{\theta})+\tilde{p}'(s;\boldsymbol{\theta})(t-s)$ for $t,s\geq 0$ as in Chapter 4. Then, for $\boldsymbol{\alpha}_r=\boldsymbol{\beta}^{(k)}$

$$\begin{aligned}\xi(\boldsymbol{\phi}) &\leq g_N(\boldsymbol{\phi})+D_N(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})+J_N(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}^{(k)})\\ &\quad+\sum_{r=1}^{K}\pi_r^{\upsilon}\left\{q(\boldsymbol{\beta}_r,\boldsymbol{\beta}_r^{(k)};\boldsymbol{\lambda}_r)+\lambda\varepsilon\|\boldsymbol{\beta}_r\|^2\right\}\\ &= \xi(\boldsymbol{\phi})+D_N(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})+J_N(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}^{(k)})+R(\boldsymbol{\beta},\boldsymbol{\beta}^{(k)}),\end{aligned} \tag{5.6}$$

where $R(\boldsymbol{\beta},\boldsymbol{\beta}^{(k)})=\sum_{r=1}^{K}\pi_r^{\upsilon}\{q(\boldsymbol{\beta}_r,\boldsymbol{\beta}_r^{(k)};\boldsymbol{\lambda}_r)-p(\boldsymbol{\beta}_r;\boldsymbol{\lambda}_r)\}$.

Thus, the majorizing surrogate function in (5.6) can be expressed in the form of

$$\xi^{SUR}(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})=\xi(\boldsymbol{\phi})+\psi(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)}), \tag{5.7}$$

where $\psi(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})=D_N(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})+J_N(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}^{(k)})+R(\boldsymbol{\beta},\boldsymbol{\beta}^{(k)})$ which is nonnegative, and equal to zero only when $\boldsymbol{\phi}=\boldsymbol{\phi}^{(k)}$. Note that minimizing (5.7) in $\boldsymbol{\phi}$ is equivalent to minimizing

$$Q_N(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})+J_N(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}^{(k)})+\sum_{r=1}^{K}\pi_r^{\upsilon}\sum_{j=1}^{p}\{\tilde{q}(|\beta_{rj}|,|\beta_{rj}^{(k)}|;\boldsymbol{\lambda}_{rj})+\lambda\varepsilon\beta_{rj}^2\},$$

where $Q_N(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})=N^{-1}Q(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})$. The minimization problem simplifies further if $\upsilon=0$:

$$Q_N(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)})+J_N(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}^{(k)})+\sum_{r=1}^{K}\sum_{j=1}^{p}\{\tilde{p}'(|\beta_{rj}^{(k)}|;\boldsymbol{\lambda}_{rj})|\beta_{rj}|+\lambda\varepsilon\beta_{rj}^2\}.$$

5.2.1 General Algorithm: MIST-MIX

With an appropriate choice of $J(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$, the parameters in $\widetilde{\boldsymbol{\beta}}$ will decouple across both the components r and coefficients j. However, we can not use the Simplified MIST algorithm directly as there are other parameters in $\boldsymbol{\phi}$ that have not necessarily decoupled. We use BCD (e.g. Tseng, 2001) to get around this issue. In particular, when we want to emphasize that $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$ is a function of $\boldsymbol{\phi}$ for fixed $\boldsymbol{\phi}^{(k)}$, we also write $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$. Partition $\boldsymbol{\phi}$ into D blocks $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_D\}$, where $\boldsymbol{\Phi} = \boldsymbol{\Phi}_1 \times \boldsymbol{\Phi}_2 \times \ldots \times \boldsymbol{\Phi}_D$. The idea behind BCD is to minimize $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$ in $\boldsymbol{\phi}_d$, $d = 1, \ldots, D$, treating parameters in other blocks as constant. Under the cyclic version of BCD, the $(k+1)^{st}$ iterate is obtained as follows. Given current iterate $\boldsymbol{\phi}^{(k)}$,

$$\begin{aligned}
\boldsymbol{\phi}_1^{(k+1)} &= \text{argmin}_{\boldsymbol{\phi}_1 \in \boldsymbol{\Phi}_1} \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2^{(k)}, \ldots, \boldsymbol{\phi}_D^{(k)}) \\
\boldsymbol{\phi}_2^{(k+1)} &= \text{argmin}_{\boldsymbol{\phi}_2 \in \boldsymbol{\Phi}_2} \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}_1^{(k+1)}, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3^{(k)}, \ldots, \boldsymbol{\phi}_D^{(k)}) \\
&\ldots \\
\boldsymbol{\phi}_D^{(k+1)} &= \text{argmin}_{\boldsymbol{\phi}_D \in \boldsymbol{\Phi}_D} \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}_1^{(k+1)}, \ldots, \boldsymbol{\phi}_{D-1}^{(k+1)}, \boldsymbol{\phi}_D)
\end{aligned} \tag{5.8}$$

Here, we treat $\boldsymbol{\pi}$ and $\widetilde{\boldsymbol{\beta}}_r$, $r = 1, \ldots, K$, and possibly other parameters contained in $\boldsymbol{\phi}$, as blocks. Under BCD, with a 'separating' $J(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$, the Simplified MIST algorithm can be used to estimate $\widetilde{\boldsymbol{\beta}}_r$ for each component separately. For example, analogous to Chapter 4, we may use

$$h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) = \ell_r(\widetilde{\boldsymbol{\beta}}_r) - \ell_r(\widetilde{\boldsymbol{\alpha}}_r) - \nabla \ell_r(\widetilde{\boldsymbol{\alpha}}_r)^T(\widetilde{\boldsymbol{\beta}}_r - \widetilde{\boldsymbol{\alpha}}_r) + \varpi_r^{-1}\|\widetilde{\boldsymbol{\beta}}_r - \widetilde{\boldsymbol{\alpha}}_r\|^2 \tag{5.9}$$

for each r, with $\ell_r(\cdot) := \sum_{i=1}^N \delta_{ir}^{(k)} \ell_i(\cdot)$, $\delta_{ir}^{(k)}$ as in (5.3), and $\ell_i = \log f(y_i, \tilde{\mathbf{x}}_i' \widetilde{\boldsymbol{\beta}}_r)$ for each observation $i = 1 \ldots, N$. As in Chapter 4, $\varpi_r \in (0, 2/\lambda_r^*)$, where λ_r^* is the largest eigenvalue of $-\nabla^2 \ell_r(\cdot)$. In practice, $\varpi_r = \varpi \in (0, 2/\lambda^*)$ is used, where λ^* is the largest eigenvalue of $-\nabla^2 \tilde{\ell}(\cdot)$, where $\tilde{\ell}(\cdot) = \sum_{i=1}^N \ell_i(\cdot)$. This mainly affects the number of iterations required for the algorithms, as the step-size will be smaller using ϖ rather

than ϖ_r; the replacement does not otherwise effect the performance of the algorithm.

With these specifications and $\tau_{rj}^{(k)} = \tilde{p}'(|\beta_{rj}^{(k)}|; \boldsymbol{\lambda}_{rj})$, the minimization of (5.6) in $\widetilde{\boldsymbol{\beta}}$ decouples for each $\widetilde{\boldsymbol{\beta}}_r$. That is, given the current iterate $\widetilde{\boldsymbol{\beta}}_r^{(k)}$ we minimize

$$-\frac{1}{N}\nabla\ell_r(\widetilde{\boldsymbol{\beta}}_r^{(k)})^T(\widetilde{\boldsymbol{\beta}}_r - \widetilde{\boldsymbol{\beta}}_r^{(k)}) + \frac{1}{N\varpi_r}\|\widetilde{\boldsymbol{\beta}}_r - \widetilde{\boldsymbol{\beta}}_r^{(k)}\|^2 + \pi_r^{\upsilon}\sum_{j=1}^{p}(\tau_{rj}^{(k)}|\beta_{rj}| + \lambda\varepsilon\beta_{rj}^2),$$

with respect to $\widetilde{\boldsymbol{\beta}}_r$. The indicated objective function is evidently proportional to

$$\sum_{j=1}^{p}\left\{-\frac{1}{N}[\nabla\ell_r(\widetilde{\boldsymbol{\beta}}_r^{(k)})]_j\beta_{rj} + \frac{1}{N\varpi_r}(\beta_{rj}^2 - 2\beta_{rj}\beta_{rj}^{(k)}) + \pi_r^{\upsilon}\tau_{rj}^{(k)}|\beta_{rj}| + \pi_r^{\upsilon}\lambda\varepsilon\beta_{rj}^2\right\}$$
$$-\frac{1}{N}[\nabla\ell_r(\widetilde{\boldsymbol{\beta}}_r^{(k)})]_0\beta_{r0} + \frac{1}{N\varpi_r}(\beta_{r0}^2 - 2\beta_{r0}\beta_{r0}^{(k)}).$$

This is precisely the form required for the Simplified MIST algorithm in Chapter 4.

All that remains is the estimation of $\pi_r, r = 1, \ldots, K$. We take a similar approach as in Städler et al. (2010), who incorporate an "improvement update" for $\boldsymbol{\pi}$ when $\upsilon > 0$. In particular, they update by

$$\boldsymbol{\pi}^{(k+1)} = \boldsymbol{\pi}^{(k)} + t^{(k)}(\bar{\boldsymbol{\pi}} - \boldsymbol{\pi}^{(k)}), \tag{5.10}$$

where $\bar{\boldsymbol{\pi}}$ is the usual estimator of $\boldsymbol{\pi} = \frac{\sum_{i=1}^{N}\boldsymbol{\delta}_i^{(k)}}{N}$, and $t^{(k)} \in (0, 1]$, chosen in practice to be the largest value in the grid $\{0.1^m;\ m = 0, 1, 2, \ldots\}$ such that

$$-N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{K}\delta_{ir}^{(k)}\log\pi_r + \lambda\sum_{r=1}^{K}\pi_r^{\upsilon}\|\widetilde{\boldsymbol{\varphi}}_r\|_1$$

is decreased.

In the same spirit, when $\upsilon = 1$ we consider improving

$$-N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{K}\delta_{ir}^{(k)}\log\pi_r + \sum_{r=1}^{K}\pi_r^{\upsilon}\sum_{j=1}^{p}(\tilde{q}(|\beta_{rj}|, |\beta_{rj}^{(k)}|; \boldsymbol{\lambda}_{rj}) + \lambda\varepsilon\beta_{rj}^2). \tag{5.11}$$

at the current value of $\beta_{rj}^{(k)}$ by a feasible descent step using feasible point $\bar{\boldsymbol{\pi}} = \frac{\sum_{i=1}^{N}\boldsymbol{\delta}_i^{(k)}}{N}$. In particular, we use update (5.10) with $t^{(k)} \in (0, 1]$, chosen in practice to be the largest

value in the grid $\{0.1^m; m = 0, 1, 2, \ldots\}$ such that

$$-N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{K}\delta_{ir}^{(k)}\log\pi_r + \sum_{r=1}^{K}\pi_r^{\upsilon}\sum_{j=1}^{p}(\tilde{p}'(|\beta_{rj}^{(k)}|;\boldsymbol{\lambda}_{rj})|\beta_{rj}^{(k)}| + \lambda\varepsilon(\beta_{rj}^{(k)})^2) \tag{5.12}$$

is decreased. Form (5.12) was used in implementation as opposed to

$$-N^{-1}\sum_{i=1}^{N}\sum_{r=1}^{K}\delta_{ir}^{(k)}\log\pi_r + \sum_{r=1}^{K}\pi_r^{\upsilon}\sum_{j=1}^{p}(\tilde{p}(|\beta_{rj}^{(k)}|;\boldsymbol{\lambda}_{rj}) + \lambda\varepsilon(\beta_{rj}^{(k)})^2), \tag{5.13}$$

which results from setting $\beta_{rj} = \beta_{rj}^{(k)}$ in (5.11). Both (5.12) and (5.13) lead to a descent direction (the remaining portion of $\tilde{q}(|\beta_{rj}|, |\beta_{rj}^{(k)}|; \boldsymbol{\lambda}_{rj})$ not included in (5.12) is bounded and nonnegative) and result in nearly identical estimates of $\boldsymbol{\pi}$ upon convergence of the algorithm. However, the former tends to converge faster (i.e., requires less iterations) than the latter. We remark that the improvement step (regardless of form) with $\upsilon \neq 0$, as well as the block coordinate minimization (which does not necessarily result in the actual minimum) leads to a *generalized* MM algorithm rather than a true MM algorithm in which full minimization is achieved at each step of the algorithm.

The proposed algorithm is as follows:

Initialize $\boldsymbol{\phi}^{(0)} = (\boldsymbol{\pi}^{(0)}, \widetilde{\boldsymbol{\beta}}_1^{(0)}, \ldots, \widetilde{\boldsymbol{\beta}}_K^{(0)})$ and fix υ. Set $k = 0$ and iterate until convergence:

1. Majorize: Compute $Q_N(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$, or equivalently, calculate

$$\delta_{ir}^{(k)} = \frac{\pi_r^{(k)} f(y_i; \tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_r^{(k)})}{\sum_{m=1}^{K}\pi_m^{(k)} f(y_i; \tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_m^{(k)})},\ r = 1, \ldots, K,\ i = 1, \ldots, N.$$

2. Minimize and/or Improve:

   a. ($\upsilon = 0$) Minimize $\xi_{\boldsymbol{\phi}^{(k)}}^{SUR}(\boldsymbol{\phi})$ with respect to $\boldsymbol{\pi}$ such that $0 \leq \pi_r \leq 1$, $r = 1, \ldots, K$, and $\sum_{r=1}^{K}\pi_r = 1 : \boldsymbol{\pi}^{(k+1)} = \frac{\sum_{i=1}^{N}\boldsymbol{\delta}_i^{(k)}}{N}$.

   a.′ ($\upsilon = 1$) Improve $\xi_{\boldsymbol{\phi}^{(k)}}^{SUR}(\boldsymbol{\phi})$ with respect to the probability simplex

$$\{\boldsymbol{\pi} : \sum_{r=1}^{K}\pi_r = 1,\ \ 0 \leq \pi_r \leq 1,\ r = 1, \ldots, K\}$$

by a feasible descent step using feasible point $\bar{\boldsymbol{\pi}} = \frac{\sum_{i=1}^{N} \boldsymbol{\delta}_i^{(k)}}{N}$ and (5.10) where $t^{(k)} \in (0, 1]$, chosen to be the largest value in the grid $\{0.1^m;\ m = 0, 1, 2, \ldots\}$ such that (5.12) is decreased.

b. Minimize $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$ with respect to $\widetilde{\boldsymbol{\beta}}_r$:

$$\begin{aligned} \boldsymbol{\beta}_r^{(k+1)} &= \frac{S\left(\boldsymbol{\beta}_r^{(k)} + \frac{\varpi_r}{2}[\nabla \ell_r(\widetilde{\boldsymbol{\beta}}_r^{(k)})]_{\mathcal{A}}, \frac{(\pi_r^{(k)})^{\upsilon} \varpi_r N}{2} \boldsymbol{\tau}_r^{(k)}\right)}{1 + (\pi_r^{(k)})^{\upsilon} N \varpi_r \lambda \varepsilon} \\ \beta_{r0}^{(k+1)} &= \beta_{r0}^{(k)} + \frac{\varpi_r}{2}[\nabla \ell_r(\widetilde{\boldsymbol{\beta}}_r^{(k)})]_0, \end{aligned} \tag{5.14}$$

for $r = 1, \ldots, K$, $\boldsymbol{\tau}_r^{(k)} = (\tau_{r1}^{(k)}, \ldots, \tau_{rp}^{(k)})'$, and $\mathcal{A} = \{1, \ldots, p\}$.

We provide details for the most common case of a linear mixture model below.

5.2.2 Example: Linear Mixture, Unknown Common Variance

Using the results from the above section for the case of linear models with *known* variance, the components necessary for MIST are $-N^{-1}\ell_r(\widetilde{\boldsymbol{\beta}}_r) = \frac{1}{2N}\left\|\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\beta}}_r - \mathbf{y}_r\right\|^2$ with $\nabla \ell_r(\widetilde{\boldsymbol{\beta}}_r) = \widetilde{\mathbf{X}}_r^T(\mathbf{y}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\beta}}_r)$ and

$$h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) = \varpi_r^{-1}\|\widetilde{\boldsymbol{\beta}}_r - \widetilde{\boldsymbol{\alpha}}_r\|^2 - \frac{1}{2}\|\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\beta}}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\alpha}}_r\|^2, \tag{5.15}$$

where $\varpi_r \in (0, 2/\lambda_r^*)$, and λ_r^* is the largest eigenvalue of $\widetilde{\mathbf{X}}_r'\widetilde{\mathbf{X}}_r$, with $\widetilde{\mathbf{X}}_r \equiv \mathbf{W}_r^{1/2}\widetilde{\mathbf{X}}$ and $\mathbf{W}_r = diag(\boldsymbol{\delta}_r^{(k)})$. Also define $\mathbf{y}_r = \mathbf{W}_r^{1/2}\mathbf{y}$. (Note that $\widetilde{\mathbf{X}}_r$, $\mathbf{y}_r$ and $\mathbf{W}_r$ change with each MM step k, despite being suppressed in the notation).

Estimation of the variance σ^2 requires additional care. Suppose a slightly different

form of the objective function

$$\begin{aligned}\xi(\boldsymbol{\phi}) &= -N^{-1}\sum_{i=1}^{N}\left[\log\sum_{r=1}^{K}\frac{\pi_r}{\sqrt{2\pi}\sigma}\exp\frac{-(y_i-\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}_r)^2}{2\sigma^2}\right] \\ &\quad +\sum_{r=1}^{K}\pi_r^{v}\left\{p(\boldsymbol{\beta}_r/\sigma;\boldsymbol{\lambda}_r)+\lambda\varepsilon\|\boldsymbol{\beta}_r/\sigma\|^2\right\},\end{aligned}$$

with reparameterization $\widetilde{\boldsymbol{\varphi}}_r = \widetilde{\boldsymbol{\beta}}_r/\sigma$ and $\rho = \sigma^{-1}$ such that

$$\begin{aligned}\xi(\boldsymbol{\phi}) &= -N^{-1}\sum_{i=1}^{N}\left[\log\sum_{r=1}^{K}\frac{\pi_r\rho}{\sqrt{2\pi}}\exp\frac{-(\rho y_i-\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\varphi}}_r)^2}{2}\right] \\ &\quad +\sum_{r=1}^{K}\pi_r^{v}\left\{p(\boldsymbol{\varphi}_r;\boldsymbol{\lambda}_r)+\lambda\varepsilon\|\boldsymbol{\varphi}_r\|^2\right\}.\end{aligned} \tag{5.16}$$

This reparameterization is desirable as the resulting optimization problem will be convex in each block coordinate (Städler et al., 2010).

As Städler et al. (2010) considered a similar model, it is worthwhile to contrast (5.16) with their objective function of interest. They consider

$$\xi_S(\boldsymbol{\phi}) = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\sum_{r=1}^{K}\frac{\pi_r\rho_r}{\sqrt{2\pi}}\exp\frac{-(\rho_r y_i-\tilde{\mathbf{x}}_i'\boldsymbol{\varphi}_r)^2}{2}\right]+\lambda\sum_{r=1}^{K}\pi_r^{v}\sum_{j=1}^{p}w_{rj}|\varphi_{rj}|, \tag{5.17}$$

where $w_{rj}=1$ or $w_{rj}=1/|\varphi_{rj}^{(0)}|$, corresponding to the L_1 or weighted L_1 penalties for some initial value $\varphi_{rj}^{(0)}$, $r=1,\ldots,K$, $j=1\ldots,p$. Besides the use of the (weighted) L_1 penalty, the Städler et al. (2010) formulation penalizes all parameters (including the intercept, if it is included), and uses a separate ρ_r for each component r. We incorporate portions of their BCD/GEM algorithm in our own algorithm below. The $D=K+2$ block coordinates in this context are $(\boldsymbol{\pi},\widetilde{\boldsymbol{\varphi}}_1,\ldots,\widetilde{\boldsymbol{\varphi}}_K,\rho)$.

First, from the EM majorization, we obtain

$$Q_N(\boldsymbol{\phi},\boldsymbol{\phi}^{(k)}) \propto \sum_{r=1}^{K}\frac{1}{N}\sum_{i=1}^{N}\delta_{ir}^{(k)}\left\{\frac{(\rho y_i-\tilde{\mathbf{x}}_i'\widetilde{\boldsymbol{\varphi}}_r)^2}{2}-\log(\rho)-\log(\pi_r)\right\}. \tag{5.18}$$

To define $J(\widetilde{\boldsymbol{\varphi}}, \widetilde{\boldsymbol{\vartheta}})$, let $h(\widetilde{\boldsymbol{\varphi}}_r, \widetilde{\boldsymbol{\vartheta}}_r) = \varpi_r^{-1}\|\boldsymbol{\varphi}_r - \boldsymbol{\vartheta}_r\|^2 - \frac{1}{2}\|\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\vartheta}}_r\|^2$ with $\varpi_r, \widetilde{\mathbf{X}}_r, \mathbf{y}_r$ defined as in the known variance case above. With this specification, notice that for each component r,

$$\frac{1}{2}\|\rho\mathbf{y}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r\|^2 \;\leq\; \frac{1}{2}\|\rho\mathbf{y}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r\|^2 + h(\widetilde{\boldsymbol{\varphi}}_r, \widetilde{\boldsymbol{\vartheta}}_r), \tag{5.19}$$

with equality if and only if $\widetilde{\boldsymbol{\varphi}}_r = \widetilde{\boldsymbol{\vartheta}}_r$. Again, in practice, $\varpi_r = \varpi$ is used for all r so that the eigenvalues of $\widetilde{\mathbf{X}}_r'\widetilde{\mathbf{X}}_r$ do not need to be recomputed at each iteration.

Combining (5.18) and (5.19) with the penalty majorization in (3.), the surrogate function becomes

$$\begin{aligned}\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)}) \;\propto\; \sum_{r=1}^{K}\Bigg\{ & \frac{\|\rho\mathbf{y}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r\|^2 - \|\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r^{(k)}\|^2 + \frac{2}{\varpi_r}\|\widetilde{\boldsymbol{\varphi}}_r - \widetilde{\boldsymbol{\varphi}}_r^{(k)}\|^2}{2N} \\ & - \left(\frac{\log\rho}{N} + \frac{\log\pi_r}{N}\right)\sum_{i=1}^{N}\delta_{ir}^{(k)} \\ & + \sum_{j=1}^{p}\pi_r^{\upsilon}(\tilde{q}(|\varphi_{rj}|, |\varphi_{rj}^{(k)}|; \boldsymbol{\lambda}_{rj}) + \lambda\varepsilon\varphi_{rj}^2)\Bigg\}.\end{aligned} \tag{5.20}$$

The proposed algorithm for minimizing $\xi(\boldsymbol{\phi})$ in (5.16) is as follows:

Initialize $\boldsymbol{\phi}^{(0)} = (\boldsymbol{\pi}^{(0)}, \widetilde{\boldsymbol{\varphi}}_1^{(0)}, \ldots, \widetilde{\boldsymbol{\varphi}}_K^{(0)}, \rho)$ (more detail given in Section 5.2.3) and fix υ. Set $k = 0$ and iterate until convergence:

1. Majorize: Compute $Q_N(\boldsymbol{\phi}, \boldsymbol{\phi}^{(k)})$, or equivalently, calculate

$$\delta_{ir}^{(k)} = \frac{\pi_r^{(k)}\rho^{(k)}\exp\{-\frac{1}{2}(\rho^{(k)}\mathbf{y}_i - \tilde{\mathbf{x}}_i^T\widetilde{\boldsymbol{\varphi}}_r^{(k)})^2\}}{\sum_{r=1}^{K}\pi_r^{(k)}\rho^{(k)}\exp\{-\frac{1}{2}(\rho^{(k)}\mathbf{y}_i - \tilde{\mathbf{x}}_i^T\widetilde{\boldsymbol{\varphi}}_r^{(k)})^2\}},\; r = 1, \ldots, K,\; i = 1, \ldots, N.$$

2. Minimize and/or Improve:

   a. ($\upsilon = 0$) Minimize $\xi_{\boldsymbol{\phi}^{(k)}}^{SUR}(\boldsymbol{\phi})$ with respect to $\boldsymbol{\pi}$ such that $0 \leq \pi_r \leq 1$, $r = 1, \ldots, K$, and $\sum_{r=1}^{K}\pi_r = 1 : \boldsymbol{\pi}^{(k+1)} = \frac{\sum_{i=1}^{N}\boldsymbol{\delta}_i^{(k)}}{N}$.

a.′ ($\upsilon = 1$) Improve $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$ with respect to the probability simplex

$$\{\boldsymbol{\pi} : \sum_{r=1}^{K} \pi_r = 1,\ \ 0 \leq \pi_r \leq 1,\ r = 1, \ldots, K\}$$

by a feasible descent step using feasible point $\bar{\boldsymbol{\pi}} = \frac{\sum_{i=1}^{N} \boldsymbol{\delta}_i^{(k)}}{N}$ using (5.10) such that

$$-N^{-1} \sum_{i=1}^{N} \sum_{r=1}^{K} \delta_{ir}^{(k)} \log \pi_r + \sum_{r=1}^{K} \pi_r^{\upsilon} \sum_{j=1}^{p} (\tilde{p}'(|\varphi_{rj}^{(k)}|; \boldsymbol{\lambda}_{rj}) |\varphi_{rj}^{(k)}| + \lambda \varepsilon (\varphi_{rj}^{(k)})^2) \tag{5.21}$$

is decreased.

b. Minimize $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$ with respect to the remaining parameters, $\widetilde{\boldsymbol{\varphi}}_r$ and ρ :

$$\begin{aligned} \boldsymbol{\varphi}_r^{(k+1)} &= \frac{S\left(\boldsymbol{\varphi}_r^{(k)} + \frac{\varpi_r}{2}\left[\widetilde{\mathbf{X}}_r'(\rho^{(k)}\mathbf{y}_r) - \widetilde{\mathbf{X}}_r'\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r^{(k)}\right]_{\mathcal{A}}, \frac{(\pi_r^{(k+1)})^{\upsilon}\varpi_r N}{2}\boldsymbol{\tau}_r^{(k)}\right)}{1 + (\pi_r^{(k+1)})^{\upsilon} N \varpi_r \lambda \varepsilon}, \\ \varphi_{r0}^{(k+1)} &= \varphi_{r0}^{(k)} + \frac{\varpi_r}{2}\left[\widetilde{\mathbf{X}}_r'(\rho^{(k)}\mathbf{y}_r) - \widetilde{\mathbf{X}}_r'\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r^{(k)}\right]_0, \end{aligned} \tag{5.22}$$

for $r = 1, \ldots, K$ where $\boldsymbol{\tau}_r^{(k)} = (\tilde{p}'(|\varphi_{r1}^{(k)}|, \boldsymbol{\lambda}_{r1}), \ldots, \tilde{p}'(|\varphi_{rp}^{(k)}|, \boldsymbol{\lambda}_{rp}))'$, $j = 1, \ldots, p$, and $\mathcal{A} = \{1, \ldots, p\}$;

$$\rho^{(k+1)} = \frac{\sum_{r=1}^{K} \mathbf{y}_r'\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r^{(k+1)} + \sqrt{(\sum_{r=1}^{K} \mathbf{y}_r'\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r^{(k+1)})^2 + 4N(\sum_{r=1}^{K} \|\mathbf{y}_r\|^2)}}{2\sum_{r=1}^{K} \|\mathbf{y}_r\|^2}. \tag{5.23}$$

Note that all elements in $\boldsymbol{\varphi}_r$ are updated simultaneously using the soft thresholding operator, and are considered with φ_{r0} as a single block.

Remark 5.2.1. *We treat all of the elements of $\widetilde{\boldsymbol{\varphi}}_r$ as a single block. This avoids the need to cycle through each of the coefficients on the j scale, using the updated versions of the other j coefficients; i.e., we update all of the coefficients in $\boldsymbol{\varphi}_r$ at once using the soft thresholding operator, and do not have to compute each coefficient separately with the coefficients $j' < j$ updated at their new values. Städler et al. (2010) treat each individual*

parameter as a block, with the exception of the mixture parameters π_r, $r = 1, \ldots, K$ which are treated as a single block of size $K - 1$. In theory, this would require cycling through each coefficient φ_{rj} individually, but in practice they do not update their active set at every EM iteration (see page 17).

Remark 5.2.2. *Since Städler et al. (2010) use the convex LASSO penalty, they can exploit the strict convexity of their EM-majorized objective function in each univariate coordinate block to construct a convergent BCD algorithm even when $P > N$. While it is possible to achieve such convexity with other (nonconvex) penalty functions for certain values of tuning parameters for a single component (e.g., MCP, SCAD, Mazumder et al., 2009), guaranteeing strict convexity becomes much more challenging when there is more than one component. For example, consider the MCP-penalized objective function (with $\upsilon = 0$ for simplicity)*

$$g_N(\boldsymbol{\phi}) + \sum_{r=1}^{K} \sum_{j=1}^{p} \tilde{p}(|\varphi_{rj}|; \boldsymbol{\lambda}_{rj}) \tag{5.24}$$

with EM-majorized surrogate

$$g_N(\boldsymbol{\phi}) + D_N(\boldsymbol{\phi}, \boldsymbol{\theta}) + \sum_{r=1}^{K} \sum_{j=1}^{p} \tilde{p}(|\varphi_{rj}|; \boldsymbol{\lambda}_{rj}) \propto Q_N(\boldsymbol{\phi}, \boldsymbol{\theta}) + \sum_{r=1}^{K} \sum_{j=1}^{p} \tilde{p}(|\varphi_{rj}|; \boldsymbol{\lambda}_{rj}). \tag{5.25}$$

In order for (5.25) *to be strictly convex for each φ_{rj}, this requires $N^{-1}\|\widetilde{\boldsymbol{X}}_{r,j}\|^2 > 1/a$ where $\widetilde{\boldsymbol{X}}_{r,j}$ is the j^{th} column of $\widetilde{\boldsymbol{X}}_r = \boldsymbol{W}_r^{1/2}\widetilde{\boldsymbol{X}}$ with $\boldsymbol{W}_r = diag(\boldsymbol{\delta}_r^{(k)})$. As the posterior probabilities $\delta_{ir}^{(k)}$, $i = 1, \ldots, N$, $r = 1, \ldots, K$ change throughout the course of the algorithm, it is not clear how to ensure that $N^{-1}\|\widetilde{\boldsymbol{X}}_{r,j}\|^2 > 1/a$ will hold. Even with scaling the original design matrix $\widetilde{\boldsymbol{X}}$ so that $N^{-1}\sum_{i=1}^{N} x_{ij}^2 = 1$, the problem is not eliminated because of the changing posterior probabilities at each iteration. This is in contrast to the MIST-MIX algorithm which (i) uses a penalty majorization, so that the majorized version of the penalty convex in $\boldsymbol{\varphi}_r$ for each $r = 1, \ldots, K$, and (ii) uses a separation majorization to maintain simple coefficient updates, performed simultaneously for each component r using the soft-thresholding operator.*

5.2.3 Initial Values, Tuning Parameters, and Convergence Criteria

Städler et al. (2010) provide suggestions for starting values for the normal mixture case, which can similarly be used for the linear model setting presented above. In particular:

1. For each observation $i, i = 1, \ldots, N$, draw randomly a class $\kappa \in \{1, \ldots, K\}$.
2. Assign for observation i and component κ the weight $\delta_{i\kappa} = 0.90$ and weights $\delta_{ir} = 0.10/(K-1)$ to the remaining r components.
3. Set $\boldsymbol{\phi}_r^{(0)} = \mathbf{0}$, $\rho^{(0)} = 1$, and $\pi_r = 1/K$ for all $r = 1, \ldots, K$.

In terms of penalty tuning, Städler et al. (2010) use either train/validation/test data or a BIC-like criteria to choose their LASSO tuning parameter. Likewise, we minimize a modified BIC criteria

$$BIC = -2\ell(\hat{\boldsymbol{\phi}}_{K,\boldsymbol{\lambda}}) + d_e \log(n), \tag{5.26}$$

over a grid of candidates K and tuning parameters, where $\hat{\boldsymbol{\phi}}_{K,\boldsymbol{\lambda}}$ is the resulting estimator from the MIST-MIX algorithm using a maximum of K components and tuning parameters collected in $\boldsymbol{\lambda}$, and d_e is the number of non-zero parameters fit in the model. For the linear FMR model, $d_e = 1 + (G-1) + \sum_{r=1}^{G}\sum_{j=0}^{p} I(\varphi_{rj} \neq 0)$, where $G \leq K$ is the number of non-zero π_r estimates. Note that Städler et al. (2010) have $G-1$ more components to estimate, as they estimate a separate $\rho_r = 1/\sigma_r$ for each component r. We prefer estimating a single common ρ, as using separate ρ_r for each group led to somewhat unstable estimation of the number of components.

We remark that d_e is a simple approximation to actual degrees of freedom; more complicated approximations can be obtained from the trace of the approximate linear projection matrix (e.g., Zhang, 2010; Zhang et al., 2010b). However, Zhang et al.

(2010b) comment that there is little difference between the simple and more complicated estimates, and opt to use d_e in their own generalized information criterion (GIC), which encompasses AIC and BIC as special cases. Interestingly, they show that their BIC-type selector enables identification of the true model consistently, whereas their AIC-type selector tends to overfit with positive probability. These conclusions are in agreement with the features of the usual AIC and BIC in best-subset variable selection.

However, Zhang et al. (2010b) did not consider mixture models. In the (unpenalized) mixture context, it has also been observed that the AIC,

$$AIC = -2\ell(\hat{\boldsymbol{\phi}}) + 2d, \tag{5.27}$$

where d is the number of free parameters in the mixture model and $\hat{\boldsymbol{\phi}}$ is the maximum likelihood estimator, tends to overfit models and overestimate the correct number of components (e.g., McLachlan and Peel, 2000, Chapter 6, and references therein). The BIC formula in (5.26) with d_e replaced with d also has its drawbacks, despite considerable support for its use in the mixture setting (e.g., McLachlan and Peel, 2000, Chapter 6, and references therein). In some contexts (density estimation when the model for the component densities is not valid), Biernacki et al. (1998) found that that BIC also tends to fit too many components.

To help overcome the problems with BIC, Biernacki et al. (1998) introduced the Integrated Classification Likelihood Criterion (ICL), whose approximation is given by

$$-2\ell(\hat{\boldsymbol{\phi}}) - 2\sum_{r=1}^{K}\sum_{i=1}^{N}\hat{\delta}_{ir}\log\hat{\delta}_{ir} + d\log(n), \tag{5.28}$$

where $\hat{\delta}_{ir}$ represents the posterior probability for observation i belonging to group r upon convergence of the algorithm. If the components of the mixture are well-separated with posterior probabilities $\hat{\delta}_{ir}$ close to one or zero, the middle term (called the estimated entropy in McLachlan and Peel (2000)) will be close to zero. Indeed, (5.28) is equivalent

to the BIC formula when the middle term is removed. Excellent simulation performance for the ICL is reported in Biernacki et al. (1998) as well as McLachlan and Peel (2000). Thus, in addition to (5.26), we also consider minimizing a modified ICL criteria

$$ICL = -2\ell(\hat{\boldsymbol{\phi}}_{K,\boldsymbol{\lambda}}) - 2\sum_{r=1}^{K}\sum_{i=1}^{N}\hat{\delta}_{ir}\log\hat{\delta}_{ir} + d_e\log(n) \tag{5.29}$$

over a grid of candidates K and tuning parameters.

Finally, we adopt a similar convergence criteria to that in Städler et al. (2010); the algorithm is deemed to have converged if both of the following conditions hold:

$$\begin{aligned}\frac{|\xi(\boldsymbol{\phi}^{(k+1)}) - \xi(\boldsymbol{\phi}^{(k)})|}{1+|\xi(\boldsymbol{\phi}^{(k+1)})|} &\leq 10^{-5}\\ \max_{\ell}\left\{\frac{|\boldsymbol{\phi}_{\ell}^{(k+1)} - \boldsymbol{\phi}_{\ell}^{(k)}|}{1+|\boldsymbol{\phi}_{\ell}^{(k+1)}|}\right\} &\leq \sqrt{10^{-5}}.\end{aligned}$$

5.3 Convergence Results

Local convergence results for minimizing objective functions of form (5.4) are provided below for the case of $\upsilon = 0$ and fixed number of components R. More theory and a proof are provided in the Section 5.6. In order to make the presentation of these results reasonably straightforward, we impose conditions throughout that exclude the existence of stationary points that are not local minima of the functions required below.

Proposition 5.3.1. *Suppose $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, where $\boldsymbol{\Phi}$ is some convex, compact subset of $\mathbb{R}^P$ and P is the length of vector $\boldsymbol{\phi}$. Let $g_N(\boldsymbol{\phi})$ be bounded below for all $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and continuously differentiable on the interior of $\boldsymbol{\Phi}$. Assume the terms in $p(\boldsymbol{\beta}_r; \boldsymbol{\lambda}_r) = \sum_{j=1}^{p}\tilde{p}(|\beta_{rj}|; \boldsymbol{\lambda}_{rj})$ satisfy condition (P1) for all $r \in \{1,\ldots,R\}$. Let $\boldsymbol{\theta} \in \boldsymbol{\Phi}$ be an arbitrary, bounded vector with component $\boldsymbol{\alpha}$ corresponding to $\boldsymbol{\beta}$ in $\boldsymbol{\phi}$. Let $D_N(\boldsymbol{\phi}, \boldsymbol{\theta})$ be a real-valued continuous function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ that is continuously differentiable in $\boldsymbol{\phi}$ for each $\boldsymbol{\theta}$ and satisfies $D_N(\boldsymbol{\phi}, \boldsymbol{\theta}) = 0$ when $\boldsymbol{\phi} = \boldsymbol{\theta}$. Likewise, let $J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = N^{-1}\sum_{r=1}^{R} h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r)$*

where $h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) \geq 0$ *is a real-valued, continuous function of* $\widetilde{\boldsymbol{\beta}}_r$ *and* $\widetilde{\boldsymbol{\alpha}}_r$ *that is continuously differentiable in* $\widetilde{\boldsymbol{\beta}}_r$ *for each* $\widetilde{\boldsymbol{\alpha}}_r$ *and satisfies* $h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) = 0$ *when* $\widetilde{\boldsymbol{\beta}}_r = \widetilde{\boldsymbol{\alpha}}_r$ *for* $r = 1, \ldots, R$. *Let* $q(\boldsymbol{\beta}_r, \boldsymbol{\alpha}_r; \boldsymbol{\lambda}_r)$ *be defined as in* (5.5)*, with* $R(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{r=1}^{R} \{q(\boldsymbol{\beta}_r, \boldsymbol{\alpha}_r; \boldsymbol{\lambda}_r) - p(\boldsymbol{\beta}_r; \boldsymbol{\lambda}_r)\}$. *Further assume that* $g_N(\boldsymbol{\phi}) + D_N(\boldsymbol{\phi}, \boldsymbol{\theta}) + J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ *is convex in* $\boldsymbol{\phi}$ *on* $\boldsymbol{\Phi}$, *and define*

$$\psi(\boldsymbol{\phi}, \boldsymbol{\theta}) = D_N(\boldsymbol{\phi}, \boldsymbol{\theta}) + J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) + R(\boldsymbol{\beta}, \boldsymbol{\alpha}).$$

Let $\mathcal{S}$ *denote the set of stationary points of* $\xi(\cdot)$. *Assume that* $\mathcal{S}$ *is finite with at least one element; in addition, assume that each element of* $\mathcal{S}$ *corresponds to a local minimum that is interior to* $\boldsymbol{\Phi}$. *Then:*

(i) $\xi(\boldsymbol{\phi})$ *in* (5.4) *with* $\upsilon = 0$ *and fixed* R *is locally Lipschitz continuous;*

(ii) $R(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *is either identically zero or non-negative for all* $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$;

(iii) $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) \equiv \xi(\boldsymbol{\phi}) + \psi(\boldsymbol{\phi}, \boldsymbol{\theta})$ *majorizes* $\xi(\boldsymbol{\phi})$ *and the generalized MM/BCD algorithm derived from* $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ *converges to a stationary point of* $\xi(\boldsymbol{\phi})$ *if* $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ *is uniquely minimized in each block coordinate* $\boldsymbol{\phi}_d$, $d = 1, \ldots, D$, *of* $\boldsymbol{\phi}$ *for each* $\boldsymbol{\theta}$ *(using the cyclic rule); for each* $\boldsymbol{\theta}$, $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ *attains a unique minimum interior to* $\boldsymbol{\Phi}$*; at least one* $D_N(\boldsymbol{\phi}, \boldsymbol{\theta})$, $J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$, *or* $R(\boldsymbol{\beta}, \boldsymbol{\alpha})$ *is strictly positive for each* $\boldsymbol{\phi} \neq \boldsymbol{\theta}$; *and, the set of fixed points* $\mathcal{M}$ *of the generalized MM/BCD algorithm mapping* M *defined in* (5.30) *is a finite, non-empty set.*

The conditions of Proposition 5.3.1 also hold for the objective and surrogate functions respectively given in (5.16) and (5.20) with $\upsilon = 0$ and fixed number of components R; the results are provided in a corollary below with a proof in Section 5.6.

Corollary 5.3.2. *Suppose* $\boldsymbol{\phi} \in \boldsymbol{\Phi}$, *where* $\boldsymbol{\Phi}$ *is some convex, compact subset of* $\mathbb{R}^P$ *and* P *is the length of vector* $\boldsymbol{\phi}$. *Let* $\xi(\boldsymbol{\phi})$ *and* $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ *be defined as in* (5.16) *and*

(5.20)*, respectively, with* $\upsilon = 0$ *and fixed* R*. Let* $\mathcal{S}$ *denote the set of stationary points of* $\xi(\cdot)$*. Assume that* $\mathcal{S}$ *is finite with at least one element; in addition, assume that each element of* $\mathcal{S}$ *corresponds to a local minimum that is interior to* $\boldsymbol{\Phi}$*. Let* $D = R + 2$ *be the number of coordinate blocks, corresponding to* R *blocks for each* $\widetilde{\boldsymbol{\varphi}}_r$ *(each length* $p + 1$*), one block for* $\boldsymbol{\pi}$ *(of length* $R - 1$*) and one block for* ρ *(of length 1). Then the assumptions of Proposition 5.3.1 are met provided* $\widetilde{\boldsymbol{X}}$ *is* $N \times p + 1$ *where* $N > p + 1$*, and the generalized MM/BCD algorithm derived from* (5.20) *converges to a stationary point of* $\xi(\boldsymbol{\phi})$ *for* $\upsilon = 0$ *and fixed* R*.*

5.4 Simulation Results

In simulation, the performance of the MIST-MIX algorithm (for both $\upsilon = 0$ and $\upsilon = 1$) was compared using the LASSO and MCP penalties for various linear regression mixture models with unknown common variance. For simplicity, the penalization methods are referred to as LAS-0, LAS-1, MCP-0, and MCP-1 throughout this section, where the 0/1 indicates the value of υ. The various models are summarized in Table 5.1. Models M1-M4 were designed to compare balanced to unbalanced cases (in terms of sparsity and component probabilities) when the true number of components is two. Models M5 and M6 consider three-component mixtures, with balanced and unbalanced component probabilities only. As the primary interest is in sparsity, models M1-M6 were generated and fit without an intercept. In each case, models were selected to minimize (5.26) and (5.29) for a range of $K \in \{1, \ldots, 8\}$ and λ in the range of $[0.005, 0.50]$ with the parameter a fixed for MCP at 3.7. The range for λ worked well for the noise level $\sigma = 0.5$ that was used for all data-generating models. In all cases, the design matrix $\mathbf{X}$ was generated from a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$ having elements $\boldsymbol{\Sigma}_{s,t} = .5^{|s-t|}$, $1 \leq s, t \leq p$, with $p = 45$.

Table 5.1: Simulation Models.

	M1	M2	M3	M4 (a/b)	M5	M6
n	100	100	150	150	225	300
$\boldsymbol{\beta}_1$	$(0_5, 3_5, 0_{35})$	$\sim N(2, .5^2)$	$(0_5, 3_5, 0_{35})$	$\sim N(2, .5^2)$	$(3_5, 0_{40})$	$(3_5, 0_{40})$
$\boldsymbol{\beta}_2$	$(-1_5, 0_{40})$	$(-1_5, 0_{40})$	$(-1_5, 0_{40})$	$(-1_5, 0_{40})$	$(0_5, -2_5, 0_{35})$	$(0_5, -2_5, 0_{35})$
$\boldsymbol{\beta}_3$	–	–	–	–	$(0_{10}, 3_2, -2_2, 3_1, 0_{30})$	$(0_{10}, 3_2, -2_2, 3_1, 0_{30})$
$\boldsymbol{\pi}$	(.5,.5)	(.5,.5)	(.3,.7)	(.3,.7)/(.7,.3)	(1/3,1/3,1/3)	(.5,.3,.2)

Table 5.2 and 5.3 summarize the model selection results using the modified BIC and ICL criterion, respectively, including the proportion of the $B = 100$ simulations where the number of model components (two or three) was correctly (under, over) selected. For the datasets in which the number of model components was correctly selected, the proportions of correctly (under, over) identified zero and non-zero regression coefficients are also listed, with the average number of incorrect zero/nonzero classifications provided parenthetically. The final column contains the "median error" as a measure of predictive value. To compute this, for each dataset b, $C = 100$ new $\mathbf{y}$ observations were generated from the true model; the "error", or negative loglikelihood, was then computed using the selected parameter estimates associated with dataset b and the same (true) design matrix $\mathbf{X}$ for the $C = 100$ new $\mathbf{y}$ observations. The median was taken over all $B \times C$ negative loglikelihood values. Additional results are displayed in Figures 5.1 and 5.2 (modified BIC) and Figures 5.3 and 5.4 (modified ICL), which show the distribution of mixing parameter estimates (π_1 in Figures 5.1 and 5.3, and π_1 and π_2 in Figures 5.2 and 5.4).

Many of the conclusions are specific to the simulation model, but there are a few general observations that can be made about the simulation as a whole. First, across all simulations, the estimates for $\boldsymbol{\pi}$ were most variable using the LAS-1 penalization. Except for perhaps model M6, the rest of the methods estimate $\boldsymbol{\pi}$ with similar, smaller variability, although not necessarily with the same amount of bias. However, the estimates of $\boldsymbol{\pi}$ are nearly identical for the modified BIC- and ICL- selected models. One

Table 5.2: M1-M6 simulation results for modified BIC selection criterion.

		Model Selection			Coefficient Selection			Median
Model	Method	Under	Exact	Over	Under (Avg)	Exact	Over (Avg)	Error
M1	LAS-0	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (10.55)	178.73
	LAS-1	0.00	0.99	0.01	0.00 (0.00)	0.00	1.00 (19.06)	182.23
	MCP-0	0.00	0.97	0.03	0.00 (0.00)	0.25	0.75 (21.52)	179.60
	MCP-1	0.00	0.97	0.03	0.01 (0.01)	0.25	0.75 (24.69)	153.46
M2	LAS-0	0.01	0.96	0.03	0.22 (0.91)	0.00	0.99 (29.55)	171.68
	LAS-1	0.07	0.78	0.15	0.49 (2.50)	0.00	0.94 (26.09)	211.21
	MCP-0	0.00	0.83	0.17	0.31 (0.43)	0.00	1.00 (26.34)	534.57
	MCP-1	0.00	0.78	0.22	0.23 (0.27)	0.00	1.00 (26.51)	547.59
M3	LAS-0	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (13.39)	237.74
	LAS-1	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (11.54)	235.89
	MCP-0	0.00	1.00	0.00	0.00 (0.00)	0.34	0.66 (2.95)	201.27
	MCP-1	0.00	0.72	0.28	0.00 (0.00)	0.50	0.50 (1.08)	202.58
M4(a)	LAS-0	0.00	0.83	0.17	0.51 (0.98)	0.00	1.00 (32.36)	246.81
	LAS-1	0.00	0.97	0.03	0.27 (1.16)	0.00	0.96 (28.44)	246.16
	MCP-0	0.00	0.62	0.38	0.71 (2.68)	0.00	1.00 (18.66)	269.74
	MCP-1	0.00	0.68	0.32	0.65 (1.96)	0.00	1.00 (12.00)	253.43
M4(b)	LAS-0	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (25.68)	232.25
	LAS-1	0.01	0.94	0.05	0.00 (0.00)	0.01	0.99 (32.07)	230.83
	MCP-0	0.00	1.00	0.00	0.00 (0.00)	0.07	0.93 (3.14)	234.94
	MCP-1	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (10.38)	247.25
M5	LAS-0	0.03	0.97	0.00	0.00 (0.00)	0.00	1.00 (24.85)	486.42
	LAS-1	0.01	0.98	0.01	0.00 (0.00)	0.00	1.00 (22.97)	476.98
	MCP-0	0.01	0.99	0.00	0.00 (0.00)	0.19	0.81 (4.05)	404.53
	MCP-1	0.00	1.00	0.00	0.00 (0.00)	0.05	0.95 (4.93)	397.36
M6	LAS-0	0.00	1.00	0.00	0.06 (0.16)	0.00	1.00 (37.86)	598.58
	LAS-1	0.02	0.92	0.06	0.01 (0.01)	0.00	1.00 (30.23)	582.46
	MCP-0	0.00	0.96	0.04	0.00 (0.00)	0.23	0.77 (5.60)	492.72
	MCP-1	0.00	1.00	0.00	0.00 (0.00)	0.27	0.73 (1.68)	485.39

could argue that were fewer outlying estimates for $\boldsymbol{\pi}$ in Figure 5.4 with the modified ICL criteria than in Figure 5.2 for modified BIC, which were intentionally plotted with the same vertical axes. Second, the MCP methods were better able to identify the correct model in terms of number of components and non-zero coefficients than the LAS methods. Third, as noted in Städler et al. (2010), the cases with unbalanced groups (M3, M4, M6) generally benefited, or at least did not suffer, from setting $\upsilon = 1$. However, M4(b) and M6 (modified ICL) were exceptions. Evidently more heavily weighting the non-sparse component had adverse effects on the median error especially for MCP with $\upsilon = 1$. Similar median error behavior occurred in the balanced-group but unbalanced

sparsity case of M2, although to a lesser degree. As for M6 in Table 5.3, the median errors for LAS were both lower than in Table 5.2, yet the LAS-1 median error using the modified ICL criterion was higher for that of LAS-0. The behavior is puzzling as the ICL-based models, on average, contained many more coefficients incorrectly estimated as nonzero than the BIC-based models. Finally, the ICL-based models tended to have fewer components on average than the BIC-selected models. This often worked in favor of modified ICL selection in terms of identifying the correct number of components, but sometimes did not (see, in particular, results for models M2 and M4(a)). While generally selecting fewer components, the ICL-based models also generally tended to overfit the model in terms of coefficients as compared to the BIC-based models (compare penultimate columns in Tables 5.2 and 5.3). Specific results for each model will now be summarized in turn.

M1 (Balanced Sparsity, Two Balanced Groups): From Tables 5.2 and 5.3, there was only one case (same case in both tables) among the LASSO simulations in which the number of components was over-estimated (three components) with LAS-1. Similarly for both MCP methods, the three and two aberrant cases for each v had solutions with three distinct groups. Interestingly, while both MCP methods were able to identify the correct models in terms of components and nonzero coefficients in Tables 5.2 and 5.3, when the models were overfit, they estimated more coefficients incorrectly as nonzero on average than the LASSO methods. Also note that the coefficient under/over columns are not mutually exclusive, as it is possible to both incorrectly estimate coefficients as nonzero and zero within the same dataset. Indeed, this occurred for MCP. The median error results favor MCP-1 over the rest in Table 5.2, but not in Table 5.3, most likely due to the relatively large number of coefficients incorrectly estimated as nonzero. With the exception of LAS-1, Figures 5.1 and 5.3 indicate that the estimation of $\boldsymbol{\pi}$ across penalization methods was quite similar.

Table 5.3: M1-M6 simulation results for modified ICL selection criterion.

		Model Selection			Coefficient Selection			Median
Model	Method	Under	Exact	Over	Under (Avg)	Exact	Over (Avg)	Error
M1	LAS-0	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (20.45)	177.24
	LAS-1	0.00	0.99	0.01	0.00 (0.00)	0.00	1.00 (29.12)	181.25
	MCP-0	0.00	0.98	0.02	0.01 (0.01)	0.16	0.84 (27.43)	212.19
	MCP-1	0.00	0.98	0.02	0.01 (0.01)	0.18	0.82 (33.42)	362.36
M2	LAS-0	0.06	0.92	0.02	0.18 (0.51)	0.00	0.99 (30.30)	171.68
	LAS-1	0.16	0.71	0.13	0.42 (1.39)	0.00	0.97 (28.76)	206.42
	MCP-0	0.00	0.83	0.17	0.31 (0.45)	0.00	1.00 (26.34)	533.23
	MCP-1	0.00	0.83	0.17	0.23 (0.27)	0.00	1.00 (27.45)	545.76
M3	LAS-0	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (17.23)	235.49
	LAS-1	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (13.70)	232.47
	MCP-0	0.00	1.00	0.00	0.00 (0.00)	0.32	0.68 (3.38)	201.83
	MCP-1	0.00	0.78	0.22	0.00 (0.00)	0.47	0.53 (1.24)	201.44
M4(a)	LAS-0	0.03	0.85	0.12	0.51 (0.99)	0.00	1.00 (32.42)	247.04
	LAS-1	0.00	0.98	0.02	0.27 (1.11)	0.00	0.98 (28.99)	246.38
	MCP-0	0.00	0.68	0.32	0.71 (2.68)	0.00	1.00 (19.03)	269.51
	MCP-1	0.00	0.75	0.25	0.64 (2.04)	0.00	1.00 (12.07)	251.46
M4(b)	LAS-0	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (25.80)	232.23
	LAS-1	0.01	0.95	0.04	0.00 (0.00)	0.01	0.99 (32.09)	230.54
	MCP-0	0.00	1.00	0.00	0.00 (0.00)	0.07	0.93 (3.22)	235.28
	MCP-1	0.00	1.00	0.00	0.00 (0.00)	0.00	1.00 (10.33)	247.14
M5	LAS-0	0.03	0.97	0.00	0.00 (0.00)	0.00	1.00 (36.19)	481.32
	LAS-1	0.01	0.98	0.01	0.00 (0.00)	0.00	1.00 (33.33)	460.38
	MCP-0	0.01	0.99	0.00	0.00 (0.00)	0.07	0.93 (5.61)	408.04
	MCP-1	0.00	1.00	0.00	0.00 (0.00)	0.05	0.95 (5.08)	397.37
M6	LAS-0	0.02	0.98	0.00	0.04 (0.04)	0.00	1.00 (52.42)	532.66
	LAS-1	0.01	0.92	0.07	0.01 (0.01)	0.00	1.00 (37.32)	561.98
	MCP-0	0.00	1.00	0.00	0.00 (0.00)	0.20	0.80 (9.25)	494.59
	MCP-1	0.00	1.00	0.00	0.00 (0.00)	0.27	0.73 (1.70)	485.39

M2 (Unbalanced Sparsity, Two Balanced Groups): There were no discernible patterns in the datasets that did not select models with two components. In many cases, it was difficult to identify which components corresponded to the true components. Note that the modified ICL-selected models more often identified too few components. Evidently, correct estimation of zero/nonzero coefficients was more difficult in the presence of unbalanced sparsity. The median error results favor the LASSO methods in this situation, particularly LAS-0, whereas the estimates of π are slightly better for the MCP methods.

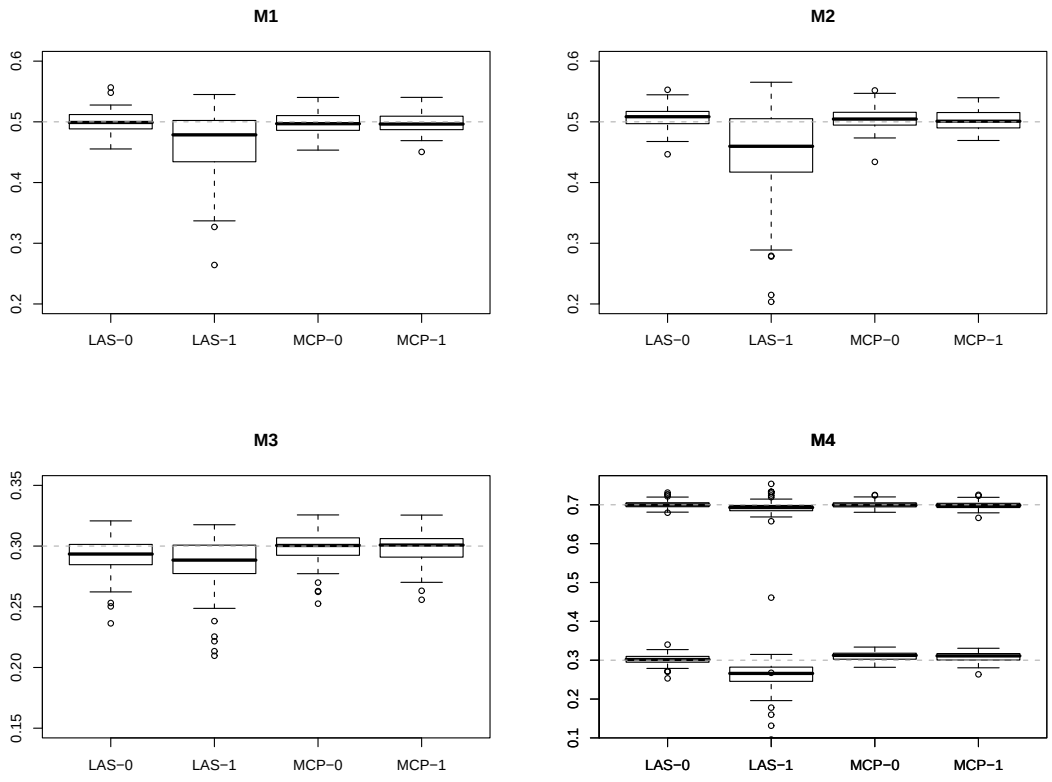

Figure 5.1: M1-M4 simulation estimates of π_1 for modified BIC-selected models.

M3 (Balanced Sparsity, Two Unbalanced Groups): From Table 5.2, MCP-1 had 28 cases in which three components were identified, all of which had solutions which split the second component into two groups, with roughly equal probabilities ($\approx .35$), with the first component estimated in nearly the correct proportion (average estimate of .296). The 22 cases from Table 5.3 were a proper subset of the 28 in Table 5.2. Evidently, the median error for MCP-1 did not suffer in either table, and in fact, the median error results favor both MCP methods in this situation. Interestingly, both MCP methods resulted in fewer coefficients incorrectly estimated as nonzero, with a substantial proportion of the datasets having both the correct number of components and coefficient classification (zero/nonzero). Not surprisingly, the estimation of $\boldsymbol{\pi}$ also favored the MCP methods.

M4(a) (Unbalanced Sparsity, Two Unbalanced Groups favoring Sparsity): There

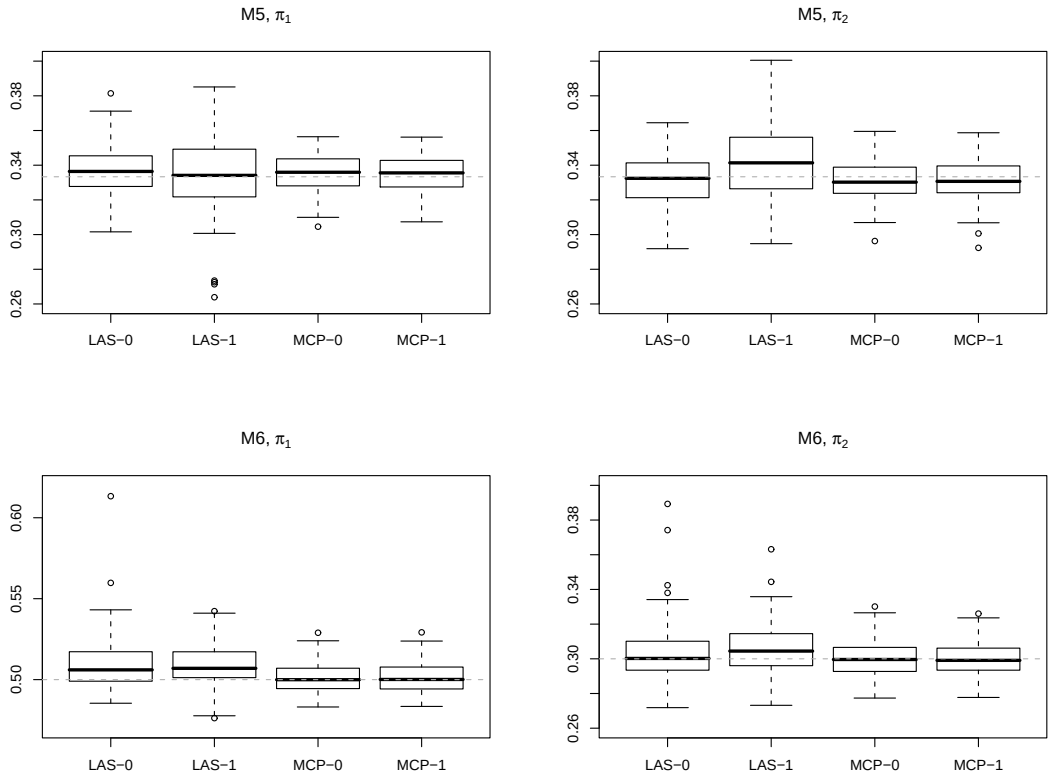

Figure 5.2: M5 and M6 simulation estimates of π_1 and π_2 for modified-BIC selected models.

were 17 cases from Table 5.2 (12 of which also appearing in Table 5.3) with LAS-0 identifying extra components; each of these cases allocated approximately 0.7 probability to a group containing predominantly negative coefficients for positions 1-5, corresponding to the second component; the first component was 'split' across two or more groups, comprising the remaining approximately 30%. Regarding LAS-1, only three datasets for the modified BIC-selected models estimated more than two components, with no discernable pattern; two of these were shared in with the modified ICL selection. Both MCP methods identified a relatively large number of datasets with more than two groups. As with LAS-0, the aberrant component datasets for MCP-0 allocated approximately 0.7 probability to a group containing predominantly negative coefficients for positions 1-5, corresponding to the second component, while the first component

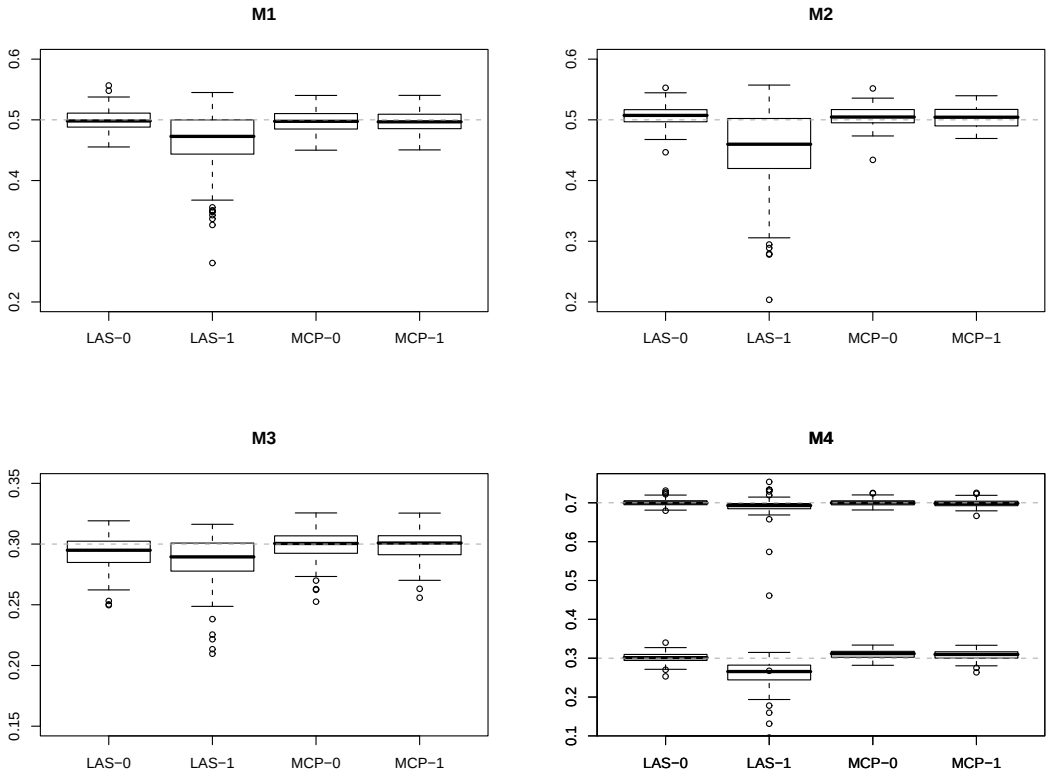

Figure 5.3: M1-M4 simulation estimates of π_1 for modified-ICL selected models.

was 'split' across two or more groups. For the aberrant component datasets using MCP-1, there were approximately the same number of datasets that 'split' the first component as those that 'split' the second component. Interestingly, both MCP methods resulted in fewer coefficients incorrectly classified, but the median error results slightly favor the LASSO methods in this situation. Estimation of $\boldsymbol{\pi}$ was also slightly better for LAS-0.

M4(b) (Unbalanced Sparsity, Two Unbalanced Groups favoring Non-Sparsity): Only LAS-1 had difficulty identifying the correct number of components in this situation. Evidently the $\upsilon = 0$ methods dominated the $\upsilon = 1$ penalizations in terms of median error, however, both MCP methods resulted in fewer incorrectly classified coefficients. The estimates for $\boldsymbol{\pi}$ were consistently good for all methods, again with the exception of LAS-1.

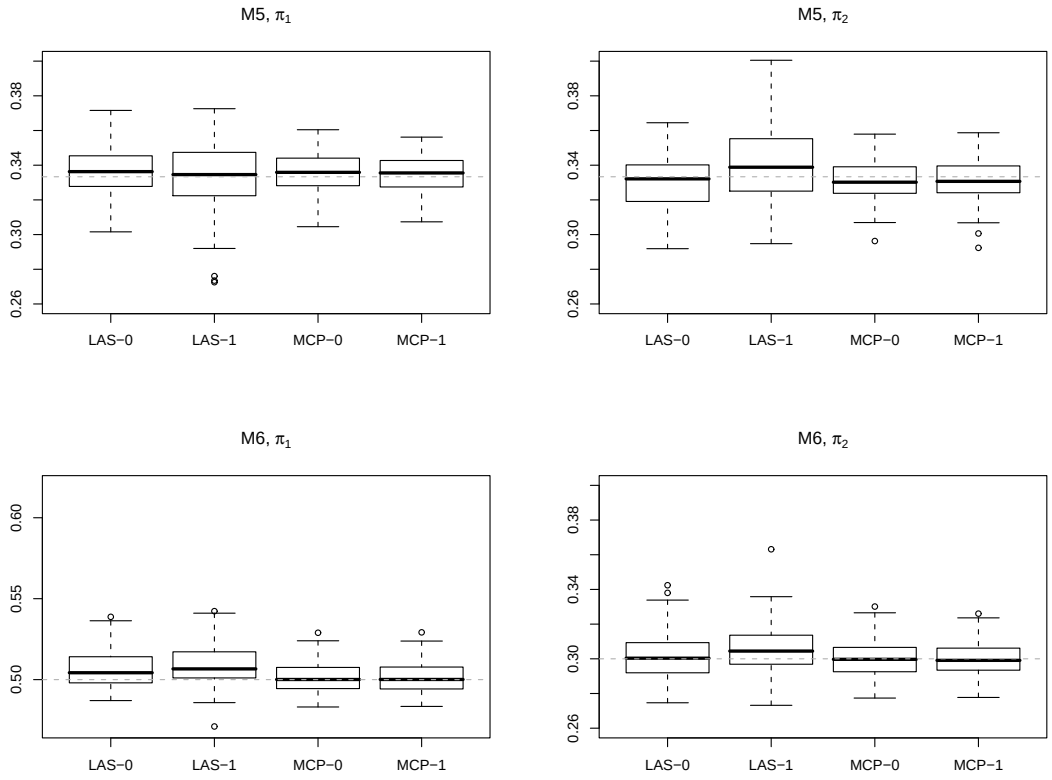

Figure 5.4: M5 and M6 simulation estimates of π_1 and π_2 for modified-ICL selected models

M5 (Balanced Sparsity, Three Balanced Groups): Both MCP methods performed better in terms of component and coefficient correctness, as well as median error, than the LASSO methods. There was very little difference in terms of modified BIC and modified ICL model selection in terms of number of components, yet the ICL-selected models tended to estimate more coefficients incorrectly as nonzero. MCP-1 seems to slightly outperform MCP-0.

M6 (Balanced Sparsity, Three Unbalanced Groups): As in M5, both MCP methods performed better in terms of component and coefficient correctness, as well as median error, than the LASSO methods; MCP-1 seems to slightly outperform MCP-0. Estimation of $\boldsymbol{\pi}$ also seems to favor the MCP methods.

Table 5.4: Ozone Data Covariates.

y :	ozone concentration	x_5 :	inversion height
x_1 :	500 mb height	x_6 :	pressure gradient
x_2 :	wind speed	x_7 :	inversion temperature
x_3 :	humidity	x_8 :	visibility
x_4 :	surface temperature	x_9 :	day of year

5.5 Example: Ozone Data

The popular ozone dataset consists of ozone concentrations and nine meteorological measurements in the Los Angeles Basin for 330 days of 1976 (see Table 5.4). The data has been previously analyzed in Breiman (1995) and Lee et al. (2006), among others. In particular, Breiman (1995) used the data to illustrate model selection with the largest model considered containing all first and second order terms. The covariates in the models derived from subset selection and his nonnegative garrote method are given in rows one and two of Table 5.5, BI and BII, respectively, along with the AIC values as provided by Lee et al. (2006) and the subsequently calculated BIC values.

Lee et al. (2006) remark that neither the BI nor BII models have well-formed polynomials; i.e., they contain product and squared terms without the corresponding main effects. Thus, models LNPI and LNPII were considered (Table 5.5, rows 3 and 4), which add the missing marginal terms to models BI and BII, respectively. Lee et al. (2006) further remark on a more serious issue regarding models BI and BII: the assumption of normality of errors with equal variance is wrong (see specifically pages 61 and 62 of Lee et al. (2006)). Thus, Lee et al. (2006) consider a GLM with gamma error and log link and use model LNPIII (Table 5.5, row 5). Based on the text description, it is unclear how they arrived at such a model; however, the use of the logarithm and squared terms make sense from simple plots of the data. The histogram of $y =$ ozone concentration is clearly skewed (Figure 5.5), and a log-transformation alleviates some of the skewness.

Table 5.5: Some existing ozone models. Most models assume a linear model structure, but LNPIII and BBW use generalized linear models with Gamma errors and a log link function.

Type	Response	Covariates Considered	Covariates Selected	(#)	AIC	BIC
BI	Normal y	1st & 2nd order, no x9	$x_6, x_2x_4, x_2x_5, x_4^2, x_6^2$	5	1934.3	1960.9
BII	Normal y	1st & 2nd order, no x9	$x_1, x_5, x_2^2, x_4^2, x_6^2, x_2x_4, x_5x_7$	7	1937.3	1971.5
LNPI	Normal y	–	$x_2, x_4, x_5, x_6, x_2x_4, x_2x_5, x_4^2, x_6^2$	8	1912.4	1950.4
LNPII	Normal y	–	$x_1, x_2, x_4, x_5, x_6, x_7,$ $x_2^2, x_4^2, x_6^2, x_2x_4, x_5x_7$	11	1913.8	1963.1
LNPIII	Gamma (log)	–	$x_2, x_4, x_7, x_8, x_9, x_8^2, x_9^2$	7	1743.3	1777.6
BBW	Gamma (log)	1st & 2nd order	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9,$ $x_3^2, x_5^2, x_6^2, x_8^2, x_9^2, x_3x_5$	15	1637.5	1698.3

Plots of $\log(y)$ $vs.$ x_j, $j = 1, \ldots, 9$ does indeed suggest the inclusion of squared terms for x_8 and x_9 (note that x_9 was not included in the Breiman (1995) analysis), although arguably a squared term for x_6 might also be appropriate (Figure 5.6).

Bar et al. (2010) recently developed an automatic variable selection and estimation procedure that is completely model-based and involves the EM algorithm. Their final model included 15 covariates, including some of the squared terms found in the LNPIII model (BBW in Table 5.5), with satisfactory residual plots. Although not by design, the model also included all first-order terms.

With such complicated models, it was surmised whether there was an underlying mixture structure that could account for some of the higher order terms. Thus, a series of models were fit using the MIST-MIX algorithm with the MCP penalty (a fixed at 3.7) and $\upsilon \in \{0, 1\}$, indicated by MMIX-0 and MMIX-1, respectively. As in the simulations, λ and K were selected by minimizing the modified BIC criterion (5.26) or the modified ICL criterion (5.29), where the range of λ was dependent on the model, but candidates K were always in the set $\{1, 2, \ldots, 15\}$. Like Breiman (1995), all variables (including interactions and squared terms) were centered and scaled prior to analysis.

As mentioned above, it was of interest to determine whether a mixture structure

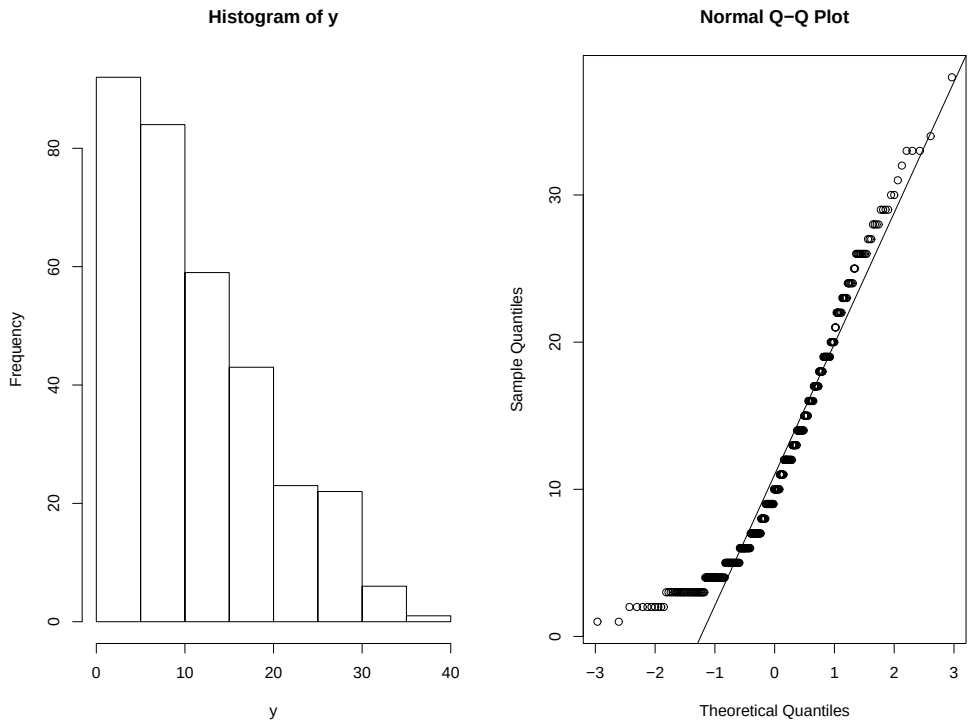

Figure 5.5: Histogram and quantile plot of response variable, ozone concentration.

with lower order terms could explain some of the higher order terms necessary in single component models. Thus, we considered fitting models with three different sets of covariates with the transformed response variable $\log(y)$: main effects only, main effects plus select second order terms, and all first and second order terms. (Unpenalized) intercepts were included in all cases, and they were allowed to differ across components. The results of the MMIX-0 and MMIX-1 analyses are in Table 5.6, as well as the usual linear model fit using the `lm` R function, denoted by LM, and the 2-component mixture model fit obtained from the R package *flexmix* (Gruen and Leisch, 2008), denoted by FMIX(2). The table is organized so that within each section, the models are nested, with the smallest appearing first and the largest (full) model appearing last. However, some comparisons can justly be made across sections (e.g., when mixture models reduce to a single component). Observe there is a dramatic decrease in the AIC and BIC values for the models in Table 5.6 as compared to Table 5.5. This is of course due to the

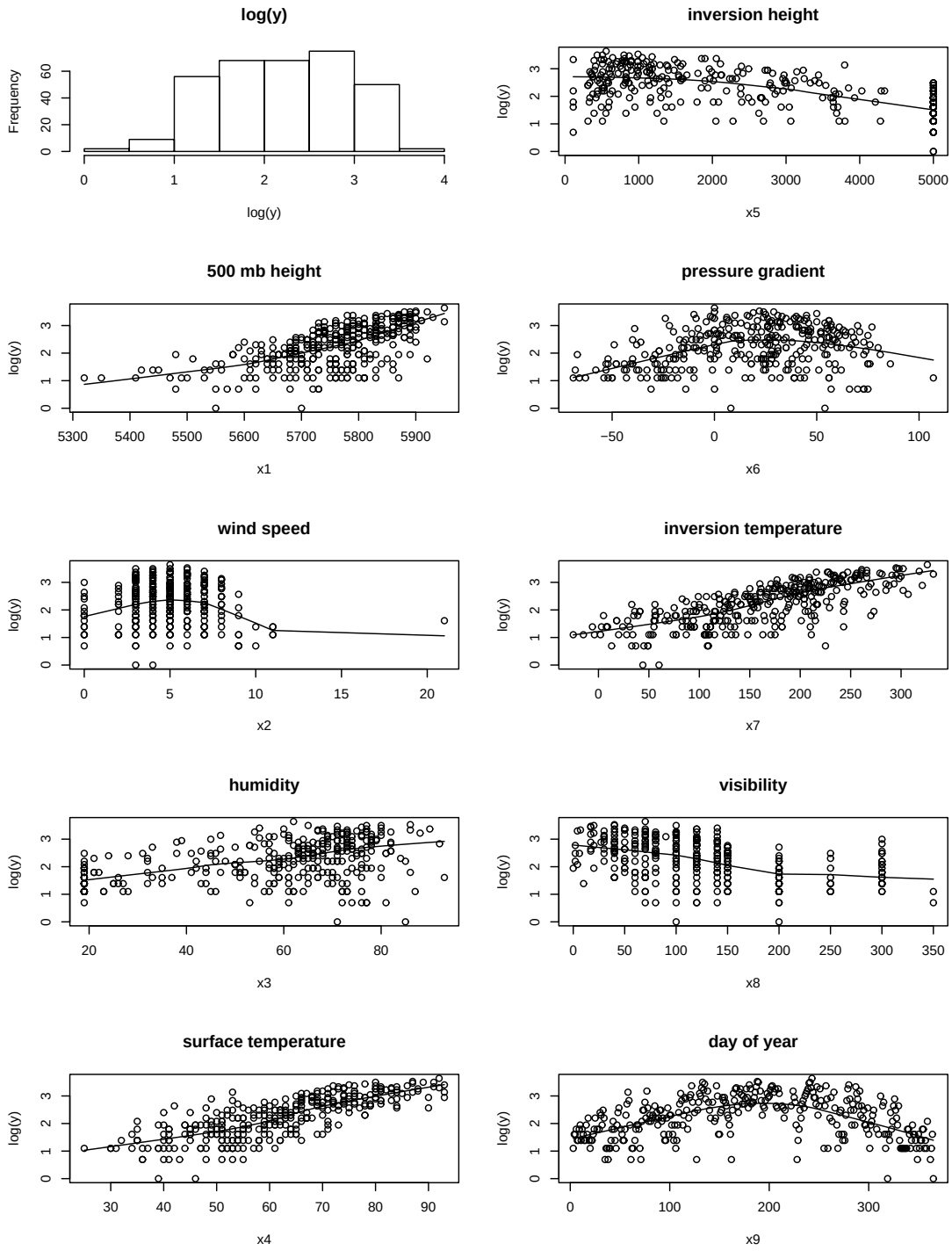

Figure 5.6: Relationships between $\log(y)$ and meteorological covariates $x_1 - x_9$, as well as the distribution of $\log(y)$.

Table 5.6: Analysis of Ozone Data, grouped by model type, all with response $\log(y)$. Model A is highlighted in light gray, whereas Model B is highlighted in a darker gray.

Covariates Considered	Covariates Selected	(#)	AIC	BIC	ICL
MMIX-0: Models selected by modified BIC					
1st order	$\hat{\boldsymbol{\beta}}_1 : x_3, x_4, x_5, x_6, x_7, x_8, x_9$ ($\hat{\pi}_1 = .82$) $\hat{\boldsymbol{\beta}}_2 : x_4, x_5, x_6, x_9$ ($\hat{\pi}_1 = .18$)	11	310.5	367.5	562.5
1st order + x_6^2, x_8^2, x_9^2	$x_1, x_2, x_4, x_5, x_6, x_8, x_9, x_6^2, x_8^2, x_9^2$	10	234.2	279.8	
1st & 2nd order	$x_2, x_4, x_6, x_7, x_8, x_9, x_1^2, x_3^2,$ $x_5^2, x_6^2, x_8^2, x_9^2, x_2x_8, x_3x_4, x_3x_5, x_6x_9$	16	**181.3**	**249.7**	
MMIX-1: Models selected by modified BIC					
1st order	$\hat{\boldsymbol{\beta}}_1 : x_3, x_4, x_5$ ($\hat{\pi}_1 = .80$) $\hat{\boldsymbol{\beta}}_2 : x_4, x_5, x_9$ ($\hat{\pi}_1 = .20$)	6	319.3	357.2	583.6
1st order + x_6^2, x_8^2, x_9^2	$\hat{\boldsymbol{\beta}}_1 : x_1, x_5, x_8, x_6^2, x_8^2, x_9^2$ ($\hat{\pi}_1 = .75$) $\hat{\boldsymbol{\beta}}_2 : x_2, x_4, x_5, x_6, x_8, x_9, x_6^2, x_9^2$ ($\hat{\pi}_1 = .25$)	14	**177.9**	**246.3**	482.0
1st & 2nd order	$x_2, x_4, x_6, x_7, x_8, x_9, x_1^2, x_3^2,$ $x_5^2, x_6^2, x_8^2, x_9^2, x_2x_8, x_3x_4, x_3x_5, x_6x_9$	16	181.4	249.7	
MMIX: Models selected by modified ICL					
1st order	x_3, x_4, x_5, x_8, x_9	5	342.7	369.3	
1st order + x_6^2, x_8^2, x_9^2	$x_1, x_2, x_4, x_5, x_6, x_8, x_9, x_6^2, x_8^2, x_9^2$	10	234.2	279.8	
1st & 2nd order	$x_2, x_4, x_6, x_7, x_8, x_9, x_1^2, x_3^2,$ $x_5^2, x_6^2, x_8^2, x_9^2, x_2x_8, x_3x_4, x_3x_5, x_6x_9$	16	**181.3**	**249.7**	
LM					
1st order	all considered	9	349.9	391.7	
1st order + x_6^2, x_8^2, x_9^2	all considered	12	237.8	**291.0**	
1st & 2nd order	all considered	54	**204.5**	417.3	
FMIX(2)					
1st order	all considered ($\hat{\boldsymbol{\pi}} = (.69, .31)$)	18	305.4	388.9	657.2
1st order + x_6^2, x_8^2, x_9^2	all considered ($\hat{\boldsymbol{\pi}} = (.76, .24)$)	24	**173.8**	**280.1**	**480.6**
1st & 2nd order	all considered ($\hat{\boldsymbol{\pi}} = (.61, .39)$)	108	173.9	599.4	827.5

loglikelihood evaluations using $\log(y)$ instead of y as a response variable, and are not directly comparable to the AIC and BIC values in Table 5.5, especially to those not using a normally distributed response.

A consistent finding throughout this analysis was that the ICL-based models always

contained one component. In models with one component, the υ specification of 0 or 1 makes no difference as $\pi = 1$. Thus, we report the modified ICL-selected models simply under the heading 'MMIX' in Table 5.6. As suggested by the simulations, selection using the modified ICL criteria often (and sometimes incorrectly) resulted in smaller models than those selected with the modified BIC criteria. We correctly anticipated the selected models with all first and second order terms to consist of a single component, but a more careful dissection is in order for the smaller models.

For MMIX-0 (modified BIC selection) and MMIX (modified ICL selection), the best model in terms of (modified) AIC and BIC was obtained as (the same) subset of covariates from the full model with a single component. For simplicity, we shall refer to this model as model A. However, the best model for MMIX-1 (modified BIC selection) in terms of (modified) AIC and BIC consists of two (unbalanced) components, and is a subset of the intermediate covariate model; call this model B. We remark here that the covariates in models A and B are not remarkably similar to any of those listed in Table 5.5. This is understandable, however, since the response variable of interest ($\log(y)$ in a linear model) is not shared among any of the existing models.

Figure 5.7 plots the distribution of the standardized residuals from model A (notably skewed), along with the partial residual plots for model A for the higher order terms not included in model B. For these partial residual plots, we color/symbol-code the residuals by group assignment according to model B. That is, each group, or component, was determined by assigning group label one to those subjects with model B posterior probabilities for group/component one greater than or equal to 0.5, and assigned to group two otherwise. Interestingly, the partial residuals for model A separate according the groups determined by model B. This suggests that the mixture in model B is indeed accounting for higher order terms present in model A.

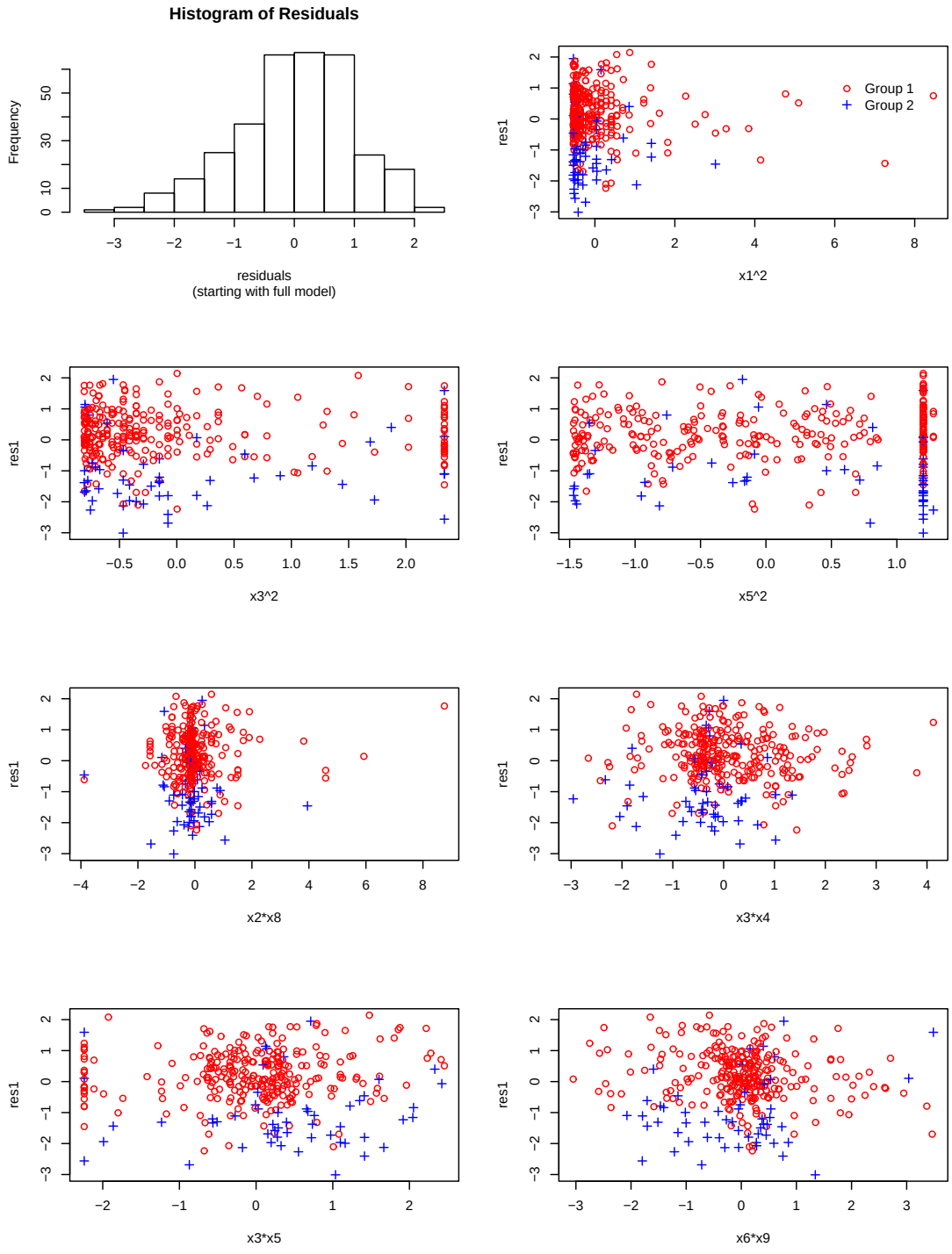

Figure 5.7: Distribution of standardized residuals for model A and partial residual plots for higher order covariates in model A not included in model B. Points are coded by most probable group status, according to model B (selected by MMIX-1 with the modified BIC criteria).

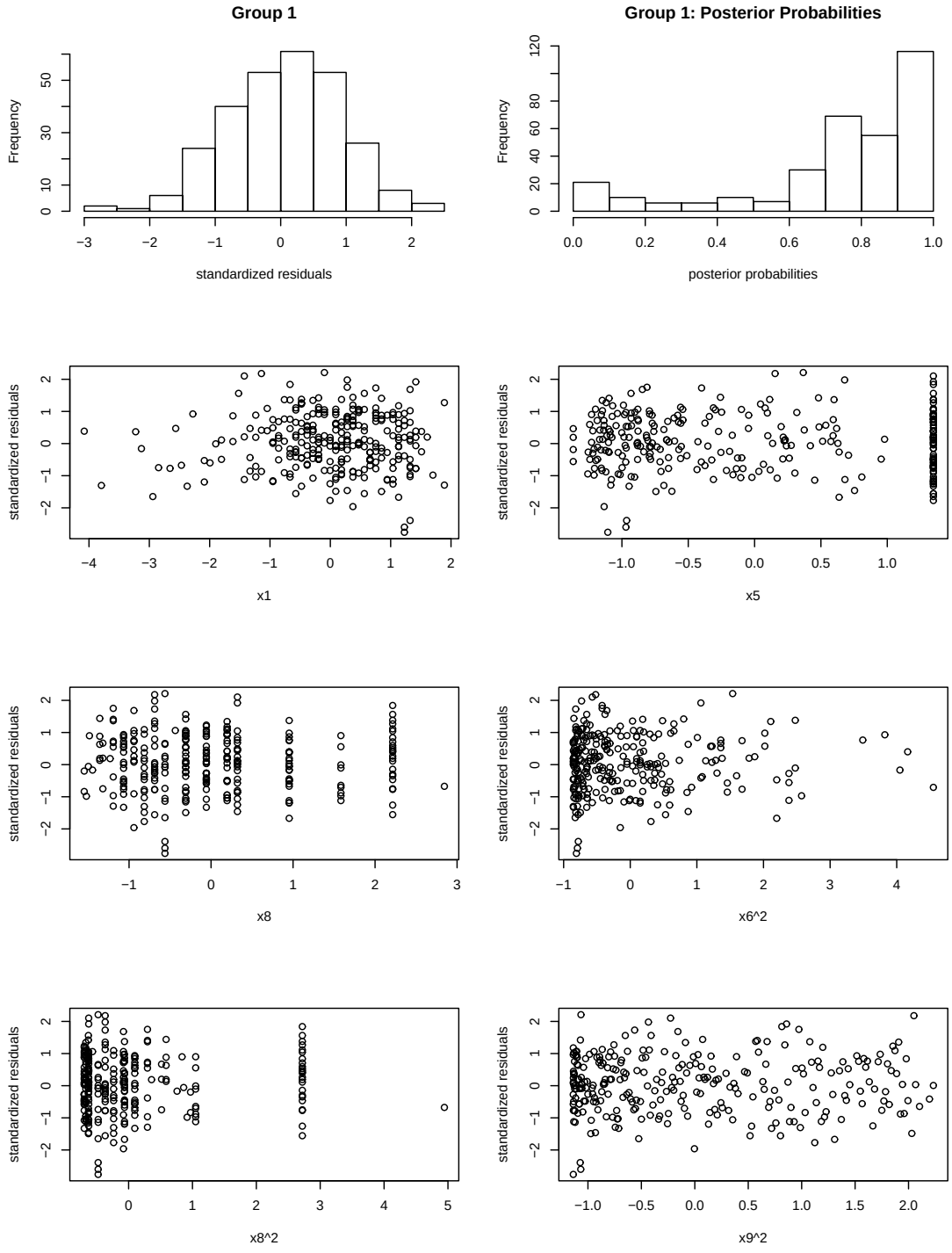

Figure 5.8: Diagnostic Plots for Group 1 in Model B: distribution of standardized residuals, distribution of posterior probabilities, and partial residual plots for relevant covariates.

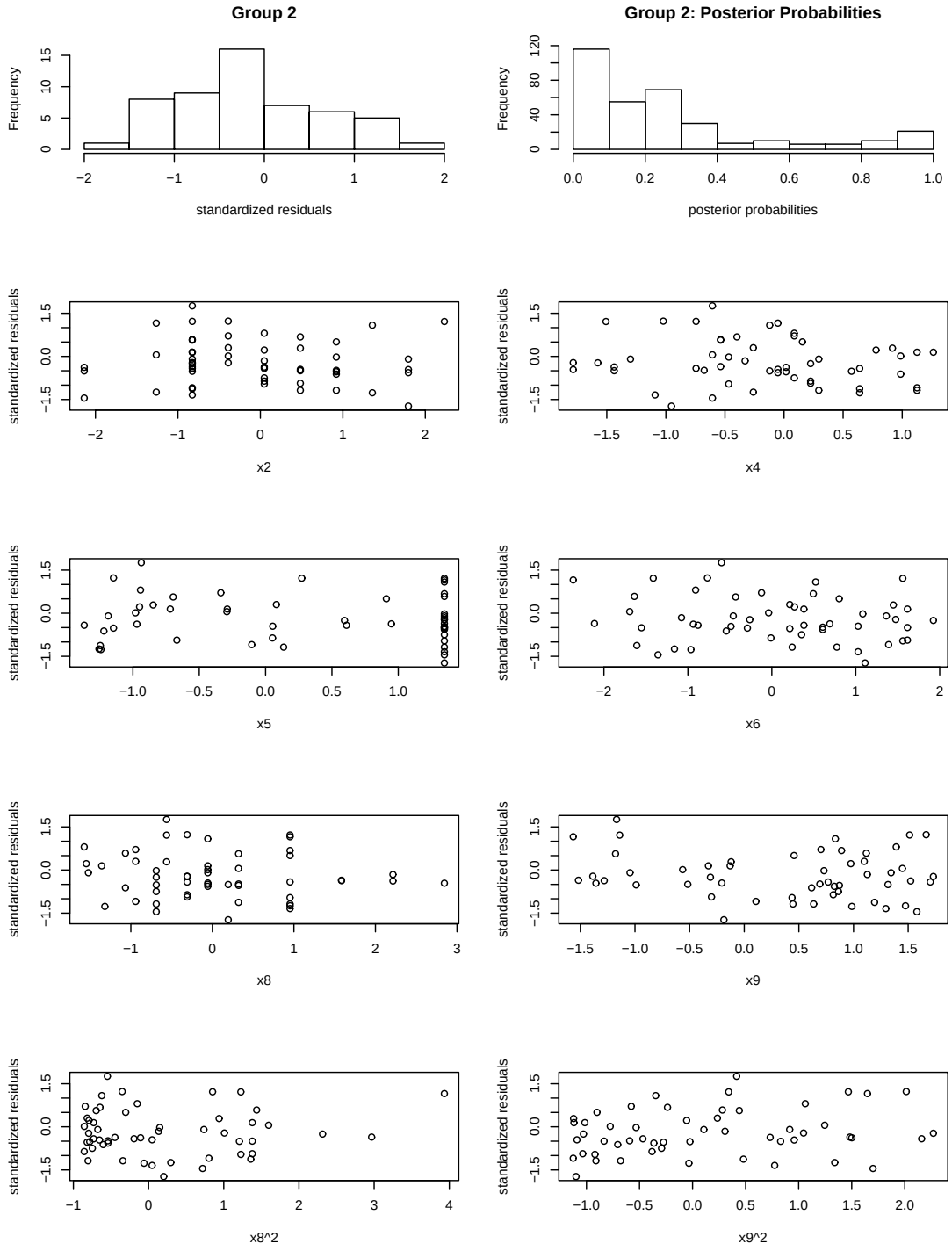

Figure 5.9: Diagnostic Plots for Group 2 in Model B: distribution of standardized residuals, distribution of posterior probabilities, and partial residual plots for relevant covariates.

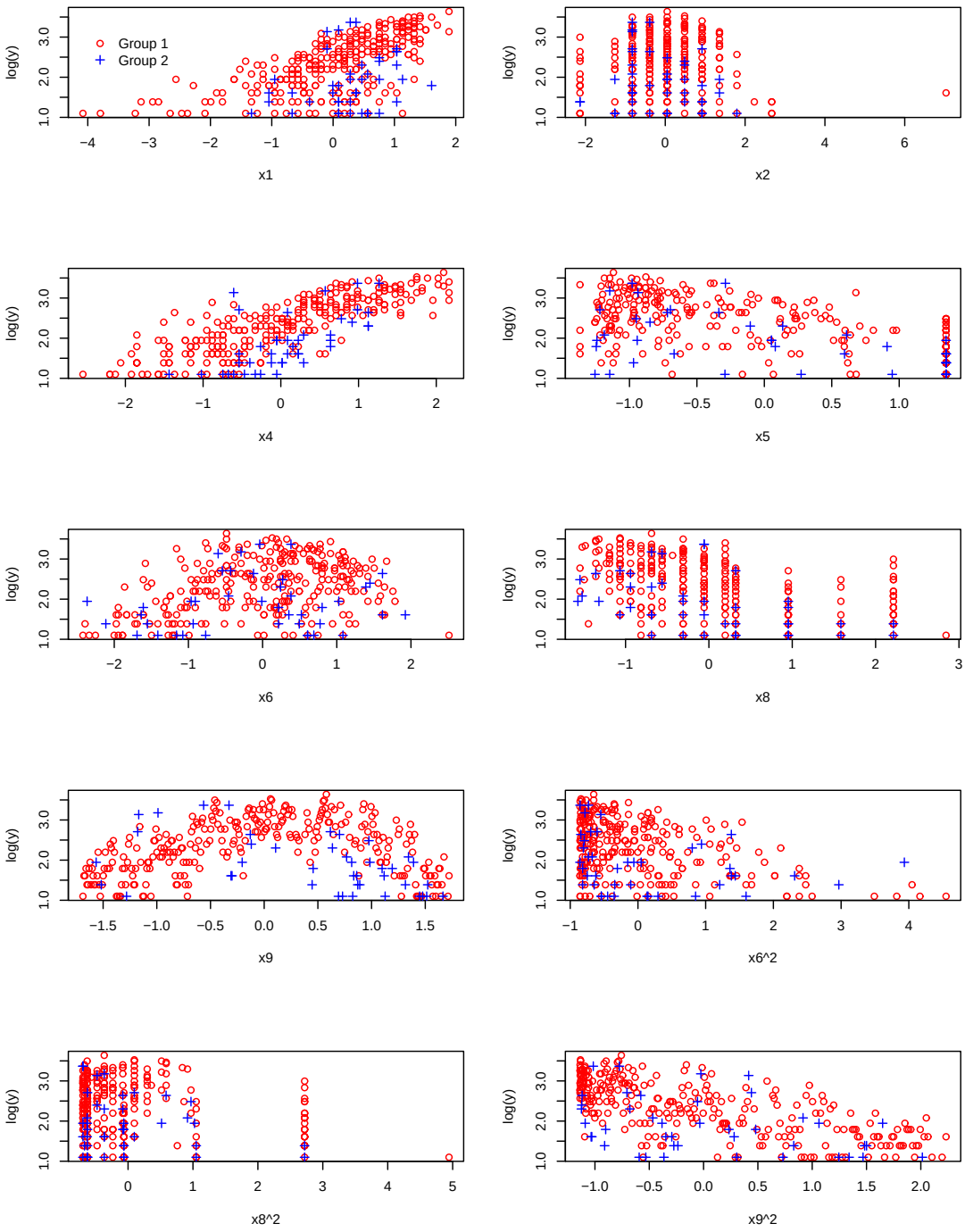

Figure 5.10: Relationships between $\log(y)$ and selected meteorological covariates, with different symbols and colors to indicate most probable group membership.

We examine the fit of model B in more detail. Figures 5.8 and 5.9 show the distribution of standardized residual distributions, as well as the distribution of the posterior probabilities and partial residual plots for the selected covariates for each component. As before, each group, or component, was determined by assigning group label one to those subjects with posterior probabilities for group/component one greater than or equal to 0.5, and assigned to group two otherwise. The histograms of the standardized residuals appear to be normally distributed, and the histograms of the posterior probabilities indicate that the groups are reasonably well-separated. The partial residual plots all appear to have random scatter, with more or less constant variance. Figure 5.10 displays $\log(y)$ plotted against each of the selected covariates, with the different colors and symbols indicating most probable group membership.

Finally, Lee, Nelder, and Pawitan might complain, however, that the MMIX-1 model does not have well-formed polynomials within each component. The *flexmix* fit with all marginal terms included for each component results in AIC=176.3 and BIC=267.4 (20 covariates, $\boldsymbol{\pi} = (.25, .75)$). While not directly comparable with model B, this new model beats the model B in terms of AIC, but not BIC; the new model can be justly compared to the full *flexmix* model, however, and beats the *flexmix* model in terms of BIC, but not AIC. In either case, a two-component model seems to provide an alternative and reasonable fit for the data.

5.6 Proofs

In this section, we provide proofs of Proposition 5.3.1 and Corollary 5.3.2. The proof of the latter involves demonstrating that the conditions of the proposition hold for the objective and surrogate functions respectively given in (5.16) and (5.20) with $\upsilon = 0$ for

fixed R. The proof of the former relies on an application of the general MM convergence of Chapter 3, with modifications in order to properly deal with our use of BCD.

In particular, we precede the proof of Proposition 5.3.1 with a proper adaptation of the theory in Chapter 3, first providing a set of sufficient conditions R1′, R2-R4, and R5′ that parallel R1-R5 given earlier and which ensure the validity of [Z1]-[Z4] (see Section 3.1) for the case of BCD applied to the MM algorithm. This will allow us to prove convergence of our BCD-based MM algorithm to a fixed point, namely a coordinatewise minimum of the proposed surrogate function. Then, we show that the conditions of Proposition 5.3.1 are sufficient to ensure R1′, R2-R4, and R5′. As shown in Meyer (1976), the set of minima of the desired objective function $\xi(\cdot)$ is a subset of the set of fixed points. As a result, we finish the proof, similarly to Städler et al. (2010), by showing that the conditions of Proposition 5.3.1 are also sufficient to ensure that each fixed point also corresponds to a minimum of the desired objective function. For this, we utilize results due to Tseng (2001) on the convergence of the BCD algorithm.

Let $\xi(\phi)$ be the real-valued function to be minimized, where $\boldsymbol{\phi} \in \boldsymbol{\Phi}$ and $\boldsymbol{\Phi}$ is some convex subset of $\mathbb{R}^P$. Let $\boldsymbol{\phi}^{(k+1)} = M(\boldsymbol{\phi}^{(k)})$, where $M : \boldsymbol{\Phi} \to \boldsymbol{\Phi}$ is the composition mapping to be defined below.

Partition $\boldsymbol{\phi}$ into D blocks such that $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_D)$. Adapting the notation of de Leeuw (1994), suppose Δ_d are D point-to-set mappings of $\boldsymbol{\Phi}$ into $\mathcal{P}(\boldsymbol{\Phi})$, the set of all subsets of $\boldsymbol{\Phi}$, such that

$$\Delta_d(\boldsymbol{\phi}) = \{\boldsymbol{\zeta} \in \boldsymbol{\Phi} : \boldsymbol{\zeta}_d = \boldsymbol{\phi}_d\}.$$

Clearly, $\boldsymbol{\phi} \in \Delta_d(\boldsymbol{\phi})$ for this choice of Δ_d for all $d = 1, \ldots, D$. Define $\Gamma_{d,\boldsymbol{\theta}}(\boldsymbol{\phi}) = \arg\min\{\xi^{SUR}_{\boldsymbol{\theta}}(\check{\boldsymbol{\phi}}) |\ \check{\boldsymbol{\phi}} \in \Delta_d(\boldsymbol{\phi})\}$ where $\xi^{SUR}_{\boldsymbol{\theta}}(\boldsymbol{\phi})$ is any function that majorizes $\xi(\boldsymbol{\phi})$ for $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. We then define $M(\boldsymbol{\phi}^{(k)}) = \Gamma_{D,\boldsymbol{\phi}^{(k)}}(\Gamma_{D-1,\boldsymbol{\phi}^{(k)}}(\ldots(\Gamma_{1,\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}^{(k)}))))$, that is,

$M(\boldsymbol{\phi}^{(k)})$ is the composition mapping such that

$$\begin{aligned} \boldsymbol{\phi}^{(k+1)[1]} &\in \Gamma_{1,\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}^{(k)}), \\ &\cdots , \\ \boldsymbol{\phi}^{(k+1)[D]} &\in \Gamma_{D,\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}^{(k+1)[D-1]}), \\ \boldsymbol{\phi}^{(k+1)} &= \boldsymbol{\phi}^{(k+1)[D]}. \end{aligned} \tag{5.30}$$

In other words, $M(\cdot)$ is a block coordinate descent mapping.

In general, $M(\cdot)$ is a point-to-set map, and therefore a set. Conditions [Z1]-[Z4] from Chapter 3 are satisfied under the following sufficient conditions, analogous to R1-R5 in Section 3.1.2:

R1′. $\xi(\boldsymbol{\phi})$ is locally Lipschitz continuous on the compact set $\boldsymbol{\Phi}$ and there exists at least one $\mathbf{z}_0 \in \boldsymbol{\Phi}$ such that $L(\xi(\mathbf{z}_0))$ is compact.

R2. $\xi(\boldsymbol{\phi}) = \xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\phi})$ for each $\boldsymbol{\phi} \in \boldsymbol{\Phi}$.

R3. $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) > \xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\phi})$ for $\boldsymbol{\phi} \neq \boldsymbol{\theta}$, $\boldsymbol{\phi}, \boldsymbol{\theta} \in \boldsymbol{\Phi}$.

R4. $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ is continuous for $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \boldsymbol{\Phi} \times \boldsymbol{\Phi}$ and locally Lipschitz continuous in $\boldsymbol{\phi}$ for $\boldsymbol{\phi}$ near $\boldsymbol{\theta}$.

R5.′ For each $\boldsymbol{\theta}, \boldsymbol{\phi} \in \boldsymbol{\Phi}$ and each $d = 1, \ldots, D$, $\Gamma_{d,\boldsymbol{\theta}}(\boldsymbol{\phi})$ is a non-empty, singleton set.

Note that condition R1′ is different from R1 as we do not assume coercivity of $\xi(\boldsymbol{\phi})$; however R1′ provides the equivalent result that the set of global minimizers of ξ on $\boldsymbol{\Phi}$ is non-empty and bounded (e.g., Ortega and Rheinboldt, 2000, Thm. 4.3.1). As before, conditions R2 and R3 imply that $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ strictly majorizes $\xi(\boldsymbol{\phi})$ and, in addition,

$$\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \xi(\boldsymbol{\phi}) + \psi(\boldsymbol{\phi}, \boldsymbol{\theta}), \tag{5.31}$$

where $\psi(\boldsymbol{\phi}, \boldsymbol{\theta}) := \xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) - \xi(\boldsymbol{\phi})$ satisfies $\psi(\boldsymbol{\phi}, \boldsymbol{\theta}) > 0$ for $\boldsymbol{\theta} \neq \boldsymbol{\phi}$ and $\psi(\boldsymbol{\phi}, \boldsymbol{\phi}) = 0$. Conditions R4 and R5′ imply that the maps $\Gamma_{d,\boldsymbol{\theta}}(\cdot)$, $d = 1, \ldots, D$, are each continuous point-to-point maps defined on compact sets, hence closed there (Zangwill, 1969, Cor. 4.2.2). This also implies the mapping M is both continuous and closed as it is the composition of closed maps on compact sets (Zangwill, 1969, Cor. 4.2.1). Conditions R1′, R4, and R5′ further imply that M always leads to a unique (coordinatewise) minimum.

Suppose R1′, R2-R4, and R5′ hold. Then, since $M(\cdot)$ is a continuous map defined on a compact set, it has at least one fixed point by Brouwer's Fixed Point Theorem (e.g., Ortega and Rheinboldt, 2000, Thm 6.3.2). Define $\mathcal{M}$ to be the set of fixed points for $M(\cdot)$. As commented above, these conditions imply that $M(\cdot)$ is closed, establishing [Z2], and condition R5′ establishes [Z4]. Propositions 5.6.1 and 5.6.2, given below and proved under conditions above, are used to establish [Z1] and [Z3′]. The conditions M1-M7 of Section 3.1 therefore hold. Among the key consequences here are the important results that all limit points of the iteration sequence are fixed points of the mapping M and the further implication that a finite set of fixed points implies convergence of the iteration sequence to one of these fixed points.

Proposition 5.6.1. *Suppose $\boldsymbol{\phi}^{(k)} \in \boldsymbol{\Phi}$ for a given $k \geq 0$. Then, $\boldsymbol{\phi}^{(k+1)} = M(\boldsymbol{\phi}^{(k)})$ exists, is bounded and is unique. In addition, for $k \geq 0$,*

$$\xi^{SUR}(\boldsymbol{\phi}^{(k+1)}, \boldsymbol{\phi}^{(k)}) \leq \xi^{SUR}(\boldsymbol{\phi}^{(k)}, \boldsymbol{\phi}^{(k)}) < \infty \tag{5.32}$$

and

$$\xi(\boldsymbol{\phi}^{(k+1)}) - \xi(\boldsymbol{\phi}^{(k)}) \leq -\psi(\boldsymbol{\phi}^{(k+1)}, \boldsymbol{\phi}^{(k)}) \leq 0. \tag{5.33}$$

where the first inequalities in both (5.32) *and* (5.33) *are strict unless $\boldsymbol{\phi}^{(k+1)} = M(\boldsymbol{\phi}^{(k)}) = \boldsymbol{\phi}^{(k)}$.*

Proposition 5.6.2. *Suppose $\boldsymbol{\phi}^{(0)} \in \boldsymbol{\Phi}$ and define $\xi^{(k)} = \xi(\boldsymbol{\phi}^{(k)})$ for $k \geq 0$. Then,*

$\{\xi^{(k)}, k \geq 0\}$ *is a bounded, monotone decreasing sequence. Moreover, the sequence* $\{\boldsymbol{\phi}^{(k)}, k \geq 0\}$ *is bounded and contained in the compact set* $L(\xi^{(0)})$.

Proof of Proposition 5.6.1: Let $\boldsymbol{\phi}^{(k)}$ be bounded, as it is contained in the compact set $\boldsymbol{\Phi}$, but otherwise arbitrary. The continuity of $M(\cdot)$, along with condition R5′, implies that $M(\boldsymbol{\phi}^{(k)})$ exists, is bounded, and is unique. Using (5.30) and condition R2, we have that

$$\begin{aligned} \infty > \xi(\boldsymbol{\phi}^{(k)}) &= \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}^{(k)}) \geq \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}^{(k+1)[1]}) \geq \dots \\ &\dots \geq \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}^{(k+1)[D-1]}) \geq \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}^{(k+1)}) = \xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(M(\boldsymbol{\phi}^{(k)}). \end{aligned} \tag{5.34}$$

Furthermore, (5.30) and R5′ imply that at least one of these inequalities is strict if we are not at fixed point. Hence, (5.32) holds.

To establish (5.33), note that (5.31), (5.32) and the definition of $\boldsymbol{\phi}^{(k+1)}$ imply

$$\xi^{SUR}(\boldsymbol{\phi}^{(k+1)}, \boldsymbol{\phi}^{(k)}) = \xi(\boldsymbol{\phi}^{(k+1)}) + \psi(\boldsymbol{\phi}^{(k+1)}, \boldsymbol{\phi}^{(k)}) < \infty.$$

Using (5.32) and the fact that $\xi^{SUR}(\boldsymbol{\phi}^{(k)}, \boldsymbol{\phi}^{(k)}) = \xi(\boldsymbol{\phi}^{(k)}) + \psi(\boldsymbol{\phi}^{(k)}, \boldsymbol{\phi}^{(k)}) = \xi(\boldsymbol{\phi}^{(k)})$, we further observe

$$\xi(\boldsymbol{\phi}^{(k+1)}) + \psi(\boldsymbol{\phi}^{(k+1)}, \boldsymbol{\phi}^{(k)}) \leq \xi(\boldsymbol{\phi}^{(k)}),$$

from which (5.33) is immediate, with strict inequality unless $\boldsymbol{\phi}^{(k+1)} = M(\boldsymbol{\phi}^{(k)}) = \boldsymbol{\phi}^{(k)}$. □

Proof of Proposition 5.6.2: Since $\boldsymbol{\phi}^{(0)} \in \boldsymbol{\Phi}$ is bounded, condition R1′ implies $\xi^{(0)}$ is bounded, $\boldsymbol{\phi}^{(0)} \in L(\xi^{(0)})$, and $L(\xi^{(0)})$ is compact. From Proposition 5.6.1 and condition R5′, we further observe that $\boldsymbol{\phi}^{(1)} = M(\boldsymbol{\phi}^{(0)})$ is bounded and satisfies $\boldsymbol{\phi}^{(1)} \in L(\xi^{(0)})$. Using condition R1′ once more, $\xi^{(1)} = \xi(\boldsymbol{\phi}^{(1)})$ is bounded and, by (5.33), satisfies $\xi^{(1)} \leq \xi^{(0)}$; thus, $L(\xi^{(1)}) \subset L(\xi^{(0)})$.

We now use induction. Let $\boldsymbol{\phi}^{(k)}$ be bounded for some $k \geq 1$ and satisfy $\xi^{(k)} \leq \xi^{(0)}$; then, $\xi^{(k)}$ is necessarily bounded and $\boldsymbol{\phi}^{(k)} \in L(\xi^{(k)}) \subset L(\xi^{(0)})$. It again follows from Proposition 5.6.1 and condition R5′ that $\boldsymbol{\phi}^{(k+1)} = M(\boldsymbol{\phi}^{(k)})$ is bounded and satisfies $\boldsymbol{\phi}^{(k+1)} \in L(\xi^{(k)})$. Hence, $\xi^{(k+1)}$ is bounded and satisfies $\xi^{(k+1)} \leq \xi^{(k)} \leq \xi^{(0)}$. Consequently, $L(\xi^{(k+1)}) \subset L(\xi^{(k)}) \subset L(\xi^{(0)})$ and $\boldsymbol{\phi}^{(k+1)} \in L(\xi^{(0)})$; it now follows that $\xi^{(k+1)} \leq \xi^{(k)}$, $L(\xi^{(k+1)}) \subset L(\xi^{(k)}) \subset L(\xi^{(0)})$, and $\boldsymbol{\phi}^{(k)} \in L(\xi^{(0)})$ for $k \geq 0$. Since $\xi(\cdot)$ is bounded below, $\{\xi^{(k)}, k \geq 0\}$ forms a bounded, monotone decreasing sequence and $\{\boldsymbol{\phi}^{(k)}, k \geq 0\}$ forms a bounded sequence contained within the compact set $L(\xi^{(0)})$. □

The results above only establish that the blockwise MM algorithm will, under certain conditions, converge to a fixed point $\bar{\boldsymbol{\phi}}$ of the mapping M, namely, a coordinatewise minimum in $\boldsymbol{\phi}$ of $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ at $\boldsymbol{\phi} = \boldsymbol{\theta} = \bar{\boldsymbol{\phi}}$. By R2, it follows that $\xi^{SUR}(\bar{\boldsymbol{\phi}}, \bar{\boldsymbol{\phi}}) = \xi(\bar{\boldsymbol{\phi}})$; however, as indicated at the beginning of this section, this does not necessarily imply that $\bar{\boldsymbol{\phi}}$ is itself a coordinatewise minimum or stationary point of $\xi(\cdot)$.

We now use the above results to prove Proposition 5.3.1. In particular, we show that the conditions of this proposition are sufficient for ensuring R1′, R2-R4, and R5′; in addition, we establish that the algorithm converges to a local minimum of ξ.

Proof of Proposition 5.3.1: The assumptions stated in the proposition immediately yield result (i), that is, $\xi(\boldsymbol{\phi})$ is locally Lipschitz continuous for each $\boldsymbol{\phi} \in \boldsymbol{\Phi}$. Result (ii) is satisfied by construction; see Chapter 4 for further details.

Recall that

$$\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \xi(\boldsymbol{\phi}) + \psi(\boldsymbol{\phi}, \boldsymbol{\theta}), \tag{5.35}$$

where

$$\psi(\boldsymbol{\phi}, \boldsymbol{\theta}) \;\; = \;\; D_N(\boldsymbol{\phi}, \boldsymbol{\theta}) + J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) + R(\boldsymbol{\beta}, \boldsymbol{\alpha}).$$

In order to establish the majorization property specified in result (iii), we begin by not-

ing that our assumptions on all components of $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ imply that $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ and $\psi(\boldsymbol{\phi}, \boldsymbol{\theta}) = \xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) - \xi(\boldsymbol{\phi})$ are both continuous in $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Our assumptions further imply that $\psi(\boldsymbol{\phi}, \boldsymbol{\theta}) \geq 0$; if at least one of its terms is strictly positive for $\boldsymbol{\phi} \neq \boldsymbol{\theta}$, then $\psi(\boldsymbol{\phi}, \boldsymbol{\theta}) > 0$ for $\boldsymbol{\phi} \neq \boldsymbol{\theta}$ and $\psi(\boldsymbol{\phi}, \boldsymbol{\phi}) = 0$. Therefore, the objective function $\xi(\boldsymbol{\phi})$ is strictly majorized by $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) \equiv \xi(\boldsymbol{\phi}) + \psi(\boldsymbol{\phi}, \boldsymbol{\theta})$.

To establish the convergence of the corresponding generalized MM algorithm in (iii), it suffices to prove that the assumptions of the proposition ensure that conditions R1′, R2-R4, and R5′ are met. Result (i), combined with the assumption that the set $\mathcal{S}$ of local minima is non-empty, finite and contained in the compact set $\boldsymbol{\Phi}$ implies the existence of at least one compact level set (Ortega and Rheinboldt, 2000, Thm 4.3.1), establishing that R1′ holds. As demonstrated in the previous paragraph, R2 and R3 also hold. By assumption, $D_N(\boldsymbol{\phi}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and continuously differentiable in $\boldsymbol{\phi}$, hence locally Lipschitz in $\boldsymbol{\phi}$. Similarly, $J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ is continuous in $\widetilde{\boldsymbol{\alpha}}$ and continuously differentiable in $\widetilde{\boldsymbol{\beta}}$, hence locally Lipschitz in $\widetilde{\boldsymbol{\beta}}$. Continuity of $q(\boldsymbol{\beta}_r, \boldsymbol{\alpha}_r; \boldsymbol{\lambda}_r) - p(\boldsymbol{\beta}_r; \boldsymbol{\lambda}_r)$ in both $\boldsymbol{\alpha}_r$ and $\boldsymbol{\beta}_r$ for all $r = 1, \ldots, R$ is also immediate. As shown Chapter 4, $q(\boldsymbol{\beta}_r, \boldsymbol{\alpha}_r; \boldsymbol{\lambda}_r) - p(\boldsymbol{\beta}_r; \boldsymbol{\lambda}_r)$ is locally Lipschitz continuous in $\boldsymbol{\beta}_r$ near $\boldsymbol{\alpha}_r$ for all r. Since both the sum and composition of two locally Lipschitz functions are locally Lipschitz, $\psi(\boldsymbol{\phi}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ and locally Lipschitz continuous in $\boldsymbol{\phi}$ near $\boldsymbol{\theta}$, implying the same for $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$. Thus, condition R4 is satisfied. Finally, condition R5′ is ensured by R1′, R2-R4, and the condition in (iii) that $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \xi^{SUR}_{\boldsymbol{\theta}}(\boldsymbol{\phi})$ in uniquely minimized in each block coordinate.

As explained earlier, these sufficient conditions imply that the mapping M defined in (5.30) has at least one fixed point $\bar{\boldsymbol{\phi}}$; in particular, $\bar{\boldsymbol{\phi}}$ is a coordinatewise minimum in $\boldsymbol{\phi}$ of $\xi^{SUR}(\boldsymbol{\phi}, \boldsymbol{\theta})$ for $\boldsymbol{\theta} = \bar{\boldsymbol{\phi}}$. Results due to Meyer (1976, pages 110-111) imply that the finite set $\mathcal{S}$ of local minimizers of $\xi(\cdot)$ is a subset of the finite set $\mathcal{M}$. It follows that

one of two situations can occur: (i) $\mathcal{S}$ and $\mathcal{M}$ are equal, hence the algorithm converges to a local minimum of ξ; or, (ii) $\mathcal{S} \subset \mathcal{M}$, in which case it is possible for the algorithm converge to a fixed point that is not a local minimum of ξ. In the latter case, it is in general possible that the coordinatewise minimum of the surrogate function at the fixed point $\boldsymbol{\theta} = \bar{\boldsymbol{\phi}}$ may not correspond to a local minimum. However, as will be shown later, the assumptions of this proposition rule out this possibility.

We first show that the coordinatewise minimum $\bar{\boldsymbol{\phi}}$ is a stationary point of $\xi_{\bar{\boldsymbol{\phi}}}^{SUR}(\boldsymbol{\phi})$ in the sense of Tseng (2001). Next, we show that stationarity in the sense of Tseng (2001) is equivalent to stationarity in the sense of Clarke (1990) for a convex function having an interior minimum. Finally, we show that the fact that $\bar{\boldsymbol{\phi}}$ is a stationary point of $\xi_{\bar{\boldsymbol{\phi}}}^{SUR}(\boldsymbol{\phi})$ in the sense of Clarke implies that it is also a stationary point of ξ under the conditions of the proposition. Since all stationary points are assumed to be local minima, it follows that the conditions of the proposition ensure convergence to a local minimum.

Tseng (2001) deals with functions f taking the form

$$f(x_1, \ldots, x_D) = f_0(x_1, \ldots, x_D) + \sum_{d=1}^{D} f_d(x_d)$$

for $f_0 : \mathbb{R}^P \rightarrow \mathbb{R} \cup \{\infty\}$ and $f_d : \mathbb{R}^{length(x_d)} \rightarrow \mathbb{R} \cup \{\infty\}$, $d = 1, \ldots, D$, with $\sum_{d=1}^{D} f_d(x_d)$ representing the separable, nondifferentiable part of f. In this context, $f = \xi_{\bar{\boldsymbol{\phi}}}^{SUR}$ and

$$\begin{aligned} f_0(\boldsymbol{\phi}) &= \xi_{\bar{\boldsymbol{\phi}}}^{SUR}(\boldsymbol{\phi}) - \sum_{r=1}^{R}\sum_{j=1}^{p} \tilde{p}'(|\bar{\beta}_{rj}|; \boldsymbol{\lambda}_{rj})|\beta_{rj}| \\ &= g_N(\boldsymbol{\phi}) + \sum_{r=1}^{R}\{\lambda\varepsilon\|\boldsymbol{\beta}_r\|^2\} + D_N(\boldsymbol{\phi}, \bar{\boldsymbol{\phi}}) + J_N(\widetilde{\boldsymbol{\beta}}, \bar{\widetilde{\boldsymbol{\beta}}}) \\ &\quad + \sum_{r=1}^{R}\sum_{j=1}^{p} \tilde{p}(|\bar{\beta}_{rj}|; \boldsymbol{\lambda}_{rj}) - \tilde{p}'(|\bar{\beta}_{rj}|; \boldsymbol{\lambda}_{rj})|\bar{\beta}_{rj}|. \end{aligned}$$

By assumption, the function $\xi_{\bar{\boldsymbol{\phi}}}^{SUR}(\boldsymbol{\phi})$ is convex and the portion we define as f_0 is

not only convex but also continuously differentiable on the interior of $\mathbf{\Phi}$. Under our assumptions, the minimum of $\xi_{\bar{\boldsymbol{\phi}}}^{SUR}(\boldsymbol{\phi})$ does not fall on the domain boundary, which is the same as the boundary of f_0. Thus, under the conditions of the proposition, condition (A2) of Tseng (2001) is necessarily satisfied and Lemma 3.1 of Tseng (2001) shows that $\bar{\boldsymbol{\phi}}$ is a stationary point $\xi_{\bar{\boldsymbol{\phi}}}^{SUR}(\boldsymbol{\phi})$ in the sense defined in that paper.

We now show that stationarity in the sense of Tseng (2001) is equivalent to stationarity in the sense of Clarke (1990) for a convex function having an interior minimum. Tseng (2001) uses the lower Dini directional derivative in order to define his notion of stationarity and in general his definition of stationary point can differ from that used in Clarke (1990). However, as we now argue, these definitions are equivalent in the case of a convex function. The following theorem is instrumental in this regard, and is a direct consequence of Theorem 1.1.1 of Hiriart-Urruty and Lemaréchal (1993, page 293) and Theorem 4.1.8 of Mäkelä and Neittaanmaäki (1992, page 64); its proof is therefore omitted.

> Theorem: let f be convex on the convex, compact set $U \subset \mathbb{R}^p$. Let x_0 be a point interior to U. Then, the following conditions are equivalent: (i) $f(x_0) \leq f(x)$ for all $x \in U$; (ii) $x_0 \in U$ satisfies $0 \in \partial f(x_0)$, where $\partial f(x_0)$ denotes the subgradient at x_0; and, (iii) $x_0 \in U$ satisfies $f'(x_0, y) \geq 0$ for all y, where $f'(x_0, y)$ denotes the directional derivative of f in direction y.

The equivalence of the definition of stationary points between Tseng (2001) and Clarke (1990) for a convex function minimized on its interior now follows directly from the above result and the fact that the subgradient of a convex function is equivalent to the notion of subdifferential defined using either the lower Dini or Clarke directional derivatives (Borwein and Lewis, 2006, Theorem 6.2.2).

The above result demonstrates that the fixed point $\bar{\boldsymbol{\phi}}$ is both a stationary point of $\xi^{SUR}_{\bar{\boldsymbol{\phi}}}(\boldsymbol{\phi})$ in the sense of Clarke (1990) and that it is the unique global interior minimum of $\xi^{SUR}_{\bar{\boldsymbol{\phi}}}(\boldsymbol{\phi})$. To finish the proof, we need only show that this stationary point is also a stationary point of $\xi(\cdot)$. Using the Clarke subdifferential (see C5 and C7 in Chapter 2) on (5.35), we have

$$\partial\xi^{SUR}_{\bar{\boldsymbol{\phi}}}(\bar{\boldsymbol{\phi}}) \subset \partial\xi(\bar{\boldsymbol{\phi}}) + \partial\psi_{\bar{\boldsymbol{\phi}}}(\bar{\boldsymbol{\phi}}).$$

Applying Remark 4.5.1 from Chapter 2, $\partial\psi_{\bar{\boldsymbol{\phi}}}(\bar{\boldsymbol{\phi}}) = \{0\}$ as $\psi_{\bar{\boldsymbol{\phi}}}(\boldsymbol{\phi})$ is minimized at $\boldsymbol{\phi} = \bar{\boldsymbol{\phi}}$. Thus, $\partial\xi^{SUR}_{\bar{\boldsymbol{\phi}}}(\bar{\boldsymbol{\phi}}) = \partial\xi(\bar{\boldsymbol{\phi}}) + \partial\psi_{\bar{\boldsymbol{\phi}}}(\bar{\boldsymbol{\phi}})$, and $\bar{\boldsymbol{\phi}}$ is also a stationary point of ξ as desired. □

Proof of Corollary 5.3.2: First observe that $g_N(\boldsymbol{\phi})$ is twice continuously differentiable, $D_N(\boldsymbol{\phi}, \boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ that is continuously differentiable in $\boldsymbol{\phi}$ for each $\boldsymbol{\theta}$ and satisfies $D_N(\boldsymbol{\phi}, \boldsymbol{\theta}) = 0$ when $\boldsymbol{\phi} = \boldsymbol{\theta}$, and $J_N(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = N^{-1}\sum_{r=1}^{R} h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r)$ where $h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) \geq 0$ is a continuous function of $\widetilde{\boldsymbol{\beta}}_r$ and $\widetilde{\boldsymbol{\alpha}}_r$ that is continuously differentiable in $\widetilde{\boldsymbol{\beta}}_r$ for each $\widetilde{\boldsymbol{\alpha}}_r$ and satisfies $h(\widetilde{\boldsymbol{\beta}}_r, \widetilde{\boldsymbol{\alpha}}_r) = 0$ when $\widetilde{\boldsymbol{\beta}}_r = \widetilde{\boldsymbol{\alpha}}_r$ for $r = 1, \ldots, R$.

Also notice that $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$ is (jointly) convex in $\boldsymbol{\phi}$: setting $\upsilon = 0$ in (5.20), we have

$$\begin{aligned}\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi}) \;\propto\; \sum_{r=1}^{R}\Bigg\{ & \frac{\|\rho\mathbf{y}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r\|^2 - \|\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r^{(k)}\|^2 + \frac{2}{\varpi_r}\|\widetilde{\boldsymbol{\varphi}}_r - \widetilde{\boldsymbol{\varphi}}_r^{(k)}\|^2}{2N} \\ & - \left(\frac{\log\rho}{N} + \frac{\log\pi_r}{N}\right)\sum_{i=1}^{N}\delta_{ir}^{(k)} \\ & + \sum_{j=1}^{p}\tilde{p}'(|\varphi_{rj}^{(k)}|;\boldsymbol{\lambda}_{rj})|\varphi_{rj}| + \lambda\varepsilon\varphi_{rj}^2)\Bigg\}.\end{aligned}$$

The only potential cause for concern is

$$\|\rho\mathbf{y}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r\|^2 - \|\widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r^{(k)}\|^2 + \frac{2}{\varpi_r}\|\widetilde{\boldsymbol{\varphi}}_r - \widetilde{\boldsymbol{\varphi}}_r^{(k)}\|^2. \tag{5.36}$$

However, $\|\rho\mathbf{y}_r - \widetilde{\mathbf{X}}_r\widetilde{\boldsymbol{\varphi}}_r\|^2$ is convex jointly in ρ and $\widetilde{\boldsymbol{\varphi}}_r$ for each r. In particular, let

$\mathbf{w} = (\rho, \widetilde{\boldsymbol{\varphi}}_r)$ so that we need only show

$$W(\mathbf{w}) = \rho^2 \|\mathbf{y}_r\|^2 - 2\rho \mathbf{y}_r' \widetilde{\mathbf{X}}_r \widetilde{\boldsymbol{\varphi}}_r + \widetilde{\boldsymbol{\varphi}}_r' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \widetilde{\boldsymbol{\varphi}}_r$$

is convex in $\mathbf{w}$. Recall that W is convex if and only if it is convex on all lines, i.e., $s(t) = W(\mathbf{w} + t\mathbf{v})$ is convex in t where $dom(s) = \{t | \mathbf{w} + t\mathbf{v} \in dom(W)\}$. Let $\mathbf{v} = (v_1, \mathbf{v}_2) \in \mathbb{R}_+ \times \mathbb{R}^{p+1}$. Then,

$$s(t) = (\rho + tv_1)^2 \|\mathbf{y}_r\|^2 - 2(\rho + tv_1)\mathbf{y}_r' \widetilde{\mathbf{X}}_r (\widetilde{\boldsymbol{\varphi}}_r + t\mathbf{v}_2) + (\widetilde{\boldsymbol{\varphi}}_r + t\mathbf{v}_2)' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r (\widetilde{\boldsymbol{\varphi}}_r + t\mathbf{v}_2)$$

and the first and second derivatives with respect to t are

$$\begin{aligned}
s'(t) &= 2v_1(\rho + tv_1)\|\mathbf{y}_r\|^2 - 2v_1 \mathbf{y}_r' \widetilde{\mathbf{X}}_r (\widetilde{\boldsymbol{\varphi}}_r + t\mathbf{v}_2) - 2(\rho + tv_1)\mathbf{y}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2 \\
&\quad + \mathbf{v}_2' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r (\widetilde{\boldsymbol{\varphi}}_r + t\mathbf{v}_2) + (\widetilde{\boldsymbol{\varphi}}_r + t\mathbf{v}_2)' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2 \\
s''(t) &= 2v_1^2 \|\mathbf{y}_r\|^2 - 4v_1 \mathbf{y}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2 + 2\mathbf{v}_2' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2 \\
&= 2\left[\|\mathbf{y}_r\|^2 (v_1 - \frac{\mathbf{y}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2}{\|\mathbf{y}_r\|^2})^2 + \mathbf{v}_2' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2 - \frac{(\mathbf{y}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2)^2}{\|\mathbf{y}_r\|^2} \right].
\end{aligned}$$

By the Cauchy-Schwarz inequality, $(\mathbf{y}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2)^2 \leq \|\mathbf{y}_r\|^2 \|\widetilde{\mathbf{X}} \mathbf{v}_2\|^2$ which implies

$$\mathbf{v}_2' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2 - \frac{(\mathbf{y}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2)^2}{\|\mathbf{y}_r\|^2} \geq \mathbf{v}_2' \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \mathbf{v}_2 - \|\widetilde{\mathbf{X}}_r \mathbf{v}_2\|^2 = 0$$

so that $s''(t) \geq 0$ and hence W is convex in $\mathbf{w} = (\rho, \widetilde{\boldsymbol{\varphi}}_r)$. Also notice that the sum of the last two terms in (5.36) is proportional to

$$\widetilde{\boldsymbol{\varphi}}_r' \left[\frac{2}{\varpi_r} \mathbf{I} - \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \right] \widetilde{\boldsymbol{\varphi}}_r - \frac{4}{\varpi_r} \widetilde{\boldsymbol{\varphi}}_r' \widetilde{\boldsymbol{\varphi}}_r^{(k)} + 2\widetilde{\boldsymbol{\varphi}}_r^{(k)'} \widetilde{\mathbf{X}}_r' \widetilde{\mathbf{X}}_r \widetilde{\boldsymbol{\varphi}}_r,$$

which is convex in $\widetilde{\boldsymbol{\varphi}}_r$ by the definition of ϖ_r for each r. Thus (5.36) is jointly convex in ρ and $\widetilde{\boldsymbol{\varphi}}_r$ for each r and $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$ is jointly convex in $\boldsymbol{\phi}$. With $D = R+2$, corresponding to R blocks for each $\widetilde{\boldsymbol{\varphi}}_r$ (each length $p+1$), one block for $\boldsymbol{\pi}$ (of length $R-1$) and one block for ρ (of length 1), $\xi^{SUR}_{\boldsymbol{\phi}^{(k)}}(\boldsymbol{\phi})$ is strictly convex in each block coordinate provided $N > p+1$, implying a unique minimum for block each coordinate. Thus, we may use Proposition 5.3.1 to obtain the desired result. □

CHAPTER 6

HIERARCHICAL MOTIVATION FOR MINIMAX CONCAVE PENALTY

The Minimax Concave Penalty (MCP) of Zhang (2010) has received substantial attention within last year among the penalized modeling community. Zhang (2010) showed that MCP, as well as the SCAD and LASSO penalties, belong to a family of quadratic spline penalties that possesses the desired sparsity and continuity properties. While both MCP and SCAD also possess the third desirable property of unbiasedness, MCP is simpler than SCAD as the former requires two knots instead of three. Simulation advantages of MCP over other penalties have been documented in both Zhang (2010) and Mazumder et al. (2009).

Critically important to any penalized optimization problem, including those using MCP, is the selection of appropriate tuning or penalty parameters. Typically, these parameters are assumed to be fixed for the purpose of estimation, and then selected a posteriori by minimizing some criteria (e.g., AIC, BIC, C_p, k-fold cross-validation, generalized cross-validation, etc.). An alternative idea, explored recently in Park and Casella (2008) and Strawderman and Wells (2010), is to take a more Bayesian approach and use a hierarchical model with a prior distribution imposed on the penalty parameter. In such Bayesian formulations, the issue of tuning parameter selection shifts to the issue of prior hyperparameter selection.

With such encouraging simulation performance of MCP, it was natural to wonder whether this penalty could be motivated from a Bayesian perspective. Among other things, such connections may provide useful insight into its good properties as well as assistance in tuning parameter selection. In a Bayesian formulation, the penalty function is regarded as the negative logarithm of the prior distribution on the coefficients or means (and possibly penalty parameters). As mentioned in Chapter 2, the LASSO estimate for

the linear regression coefficients is the maximum a posteriori (MAP) estimator when the coefficients have independent double exponential priors (e.g., Tibshirani, 1996) with parameter λ treated as known. Alternatively, the double exponential distribution can be represented hierarchically by imposing normal priors on the regression coefficients and independent exponential priors on their variances; that is, by a scale mixture of normals with an exponential mixing density (e.g., Griffin and Brown, 2005, 2007; Park and Casella, 2008).

Strawderman and Wells (2010) explore the connections between the proper and improper hierarchical priors of Strawderman (1971) and Takada (1979), and their respective proper Bayes and MAP estimators in the multivariate normal means problem ($\mathbf{Z} \sim N_p(\boldsymbol{\theta}, \mathbf{I}_p)$). Takada (1979) showed under an appropriate (improper) prior of the form $\pi(\boldsymbol{\theta}, \kappa) = \pi(\boldsymbol{\theta}|\kappa)\pi(\kappa)$ that the positive part (thresholding) James-Stein estimator

$$\hat{\boldsymbol{\theta}}_{JS+} = \left(1 - \frac{p-2}{\|\mathbf{Z}\|^2}\right)_+ \mathbf{Z},$$

with $\|\cdot\|$ denoting the usual Euclidean vector norm, can be considered a MAP estimator when the corresponding posterior is maximized *jointly* in the mean vector and κ. Likening κ to λ and in a similar spirit, Strawderman and Wells (2010) consider maximizing the posterior distribution in both the mean vector $\boldsymbol{\theta}$ and λ under a joint prior distribution $\pi(\boldsymbol{\theta}, \lambda|\alpha, \beta)$ defined hierarchically by

$$\pi(\boldsymbol{\theta}|\lambda, \alpha, \beta) \propto \lambda^p \exp\{-\lambda\|\boldsymbol{\theta}\|\}, \;\; \pi(\lambda|\alpha, \beta) \propto \lambda^{-p} \exp\{-\alpha(\lambda - \beta)^2\}, \tag{6.1}$$

where $\alpha, \beta > 0$ are hyperparameters. The MAP estimator for $(\boldsymbol{\theta}, \lambda)$ can be computed in closed-from by minimizing the objective function

$$L(\boldsymbol{\theta}, \lambda) = \|\mathbf{Z} - \boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\| + \alpha(\lambda - \beta)^2 \;\text{ for } \boldsymbol{\theta} \in \mathbb{R}^P, \lambda > 0, \tag{6.2}$$

jointly in $\boldsymbol{\theta}$ and λ. See Strawderman and Wells (2010) for more details; they refer to the resulting thresholding estimator for $\boldsymbol{\theta}$ as the Hierarchical Prior Grouped Lasso (HP-GLASSO).

The focus of this chapter is on objective function (6.2) for the case of $p = 1$, and the resulting estimator for θ obtained through the joint minimization in θ and λ. It turns out that the estimator is precisely the univariate MCP estimator proposed in Zhang (2010), which coincides with the firm thresholding estimator of Gao and Bruce (1997). Properties of the univariate MCP estimator are explored in greater detail, from both the frequentist and Bayesian perspective. Additionally, while the value of tuning parameter λ can be given in closed form, the solution depends on the hyperparameters α and β that must either be specified or selected in some fashion. We conclude this chapter with simulation results regarding hyperparameter selection.

6.1 Univariate Thresholding Estimator

Consider the simple univariate objective function

$$G(\theta, \lambda; z) = \varsigma^{-1}(\theta - z)^2 + \lambda|\theta| + \alpha(\beta - \lambda)^2, \quad \lambda, \alpha, \beta > 0, \tag{6.3}$$

where $g(\theta) := \varsigma^{-1}(\theta - z)^2$ is the data-fidelity or loss function between unknown θ and observed z, with $\varsigma > 0$ a known constant. The penalty, $\lambda|\theta| + \alpha(\beta - \lambda)^2$, is the negative logarithm of $\pi(\theta, \lambda|\alpha, \beta)$ defined hierarchically by (6.1) for the case of $p = 1$, where $\pi(\theta|\lambda, \alpha, \beta)$ reduces to $DoubExp(\lambda)$. Note that without the last term, (6.3) is simply a univariate LASSO problem (Tibshirani, 1996) for *known* λ, which is minimized in θ by the soft thresholding estimator given in (2.2) for $\varsigma = 2$. We refer to the penalty $\lambda|\theta| + \alpha(\beta - \lambda)^2$ as the Hierarchical Prior LASSO (HPLASSO) penalty.

Minimizing (6.3) simultaneously in θ and λ results in an estimator that shares the same form as both the univariate Minimax Concave Penalty (MCP) estimator of Zhang (2010) and the firm thresholding estimator of Gao and Bruce (1997).

Theorem 6.1.1. *Let $\beta > 0$ and $\alpha > \varsigma/4$ for some fixed, arbitrary $\varsigma > 0$. Then* (6.3)

is strictly convex for $(\theta, \lambda) \in \mathbb{R} \times \mathbb{R}_+$, *and has a unique minimum at* $\theta = \hat{\theta}_{HP}$ *and* $\lambda = \hat{\lambda}_{HP}$ *where*

$$\hat{\theta}_{HP} = \begin{cases} 0 & \text{if } |z| < \frac{\varsigma\beta}{2} \\ \frac{4\alpha}{4\alpha - \varsigma}\left(z - sign(z)\frac{\varsigma\beta}{2}\right) & \text{if } \frac{\varsigma\beta}{2} \leq |z| < 2\alpha\beta \\ z & \text{if } |z| \geq 2\alpha\beta \end{cases} . \tag{6.4}$$

and $\hat{\lambda}_{HP} = \left(\beta - \frac{|\hat{\theta}_{HP}|}{2\alpha}\right)_+$.

Proof: Suppose $z \neq 0$ as it is nonnull with probability one. The objective function (6.3) is continuous and bounded below for $(\theta, \lambda) \in \mathbb{R} \times \mathbb{R}_+$, so a minimum exists. Hence, strict convexity of (6.3) will guarantee a unique minimum on this set.

To show strict convexity, note that (6.3) may be rewritten as $\varsigma^{-1}z^2 - \varsigma^{-1}2z\theta + \varsigma^{-1}\theta^2 + \lambda|\theta| + \alpha(\lambda - \beta)^2$. Since the second term, $-\varsigma^{-1}2z\theta$, is convex, we need only show

$$W(\theta, \lambda) = \varsigma^{-1}\theta^2 + \lambda|\theta| + \alpha(\lambda - \beta)^2$$

is strictly convex for $\theta \in \mathbb{R}, \lambda \in \mathbb{R}_+, \alpha > \varsigma/4 > 0, \beta > 0$. Let $\mathbf{x} = (\theta, \lambda)$ and notice that $W(\theta, \lambda) = w(s_1(\mathbf{x}), s_2(\mathbf{x}))$, where $s_1(\mathbf{x}) = \|\mathbf{e}_1'\mathbf{x}\| = |\theta|$ and $s_2(\mathbf{x}) = \mathbf{e}_2'\lambda$, where $\mathbf{e}_i$, $i = 1, 2$ are the standard basis vectors. For $\mathbf{y} = (y_1, y_2) \in \mathbb{R}^2_+$,

$$\begin{aligned} w(\mathbf{y}) &= \varsigma^{-1}y_1^2 + y_1y_2 + \alpha y_2^2 + \alpha(\beta^2 - 2y_2\beta) \\ &= \frac{1}{2}(y_1 \;\; y_2)\begin{bmatrix} 2/\varsigma & 1 \\ 1 & 2\alpha \end{bmatrix}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \alpha(\beta^2 - 2y_2\beta) \end{aligned} \tag{6.5}$$

Recall that w is convex if and only if it is convex on all lines, i.e., $g(t) = w(\mathbf{y} + t\mathbf{v})$ is convex in t where $dom(g) = \{t | \mathbf{y} + t\mathbf{v} \in dom(w)\}$. Let $\mathbf{v} = (v_1, v_2) \in \mathbb{R}^2_+$. Then,

$$g(t) = \frac{(y_1 + tv_1)^2}{\varsigma} + (y_2 + tv_2)(y_1 + tv_1) + \alpha(y_2 + tv_2)^2 + \alpha(\beta^2 - 2(y_2 + tv_2)\beta)$$

The first and second derivatives of g with respect to t are

$$\begin{aligned} g'(t) &= \frac{2v_1}{\varsigma}(y_1 + tv_1) + (y_2 + tv_2)v_1 + (y_1 + tv_1)v_2 + 2\alpha v_2(y_2 + tv_2) - 2\alpha v_2\beta \\ g''(t) &= \frac{2v_1^2}{\varsigma} + 2v_1v_2 + 2\alpha v_2^2 = 2\left[\varsigma^{-1}\left(v_1 + \frac{\varsigma v_2}{2}\right)^2 + \alpha v_2^2 - \frac{\varsigma v_2^2}{4}\right]. \end{aligned}$$

The second derivative is guaranteed to be positive when $\alpha > \varsigma/4$, indicating that g is strictly convex so long as $\alpha > \varsigma/4$. Equivalently, w is strictly convex when the matrix in (6.5) is positive definite, i.e., when $\alpha > \varsigma/4$. Since w is monotonically increasing in each coordinate, and the functions $s_i(\mathbf{x}), i = 1, 2$, are each convex in $\mathbf{x}$, the composition $W(\theta, \lambda)$ is strictly convex for $(\theta, \lambda) \in \mathbb{R} \times \mathbb{R}_+$ when $\alpha > \varsigma/4$, and hence so is (6.3); the minimum of (6.3) is thus unique.

The form of the unique minimum can be obtained by finding (θ^*, λ^*) such that $\mathbf{0} \in \partial G\ (\theta^*, \lambda^*)$, where ∂G denotes the subdifferential of $G(\theta, \lambda)$ in (6.3). As (6.3) is differentiable in λ, it can be shown that $\lambda^* = \hat{\lambda}_{HP}$.

Determining the solution θ^* requires examination of three cases, depending on the value of θ^*. First consider $\theta^* = 0$; thus $\lambda^* = \beta > 0$. For $\theta^* = 0$ to be the unique minimizer, $G(0, \beta) = \varsigma^{-1}z^2$ must be less than $G(\theta, \beta) = \varsigma^{-1}(z - \theta)^2 + \beta|\theta|$ for all $\theta \neq 0$. Since $G(\theta, \beta) = \varsigma^{-1}(z^2 - 2z\theta + \theta^2) + \beta|\theta|$, we need

$$\varsigma^{-1}\theta^2 + \beta|\theta| > 2\varsigma^{-1}z\theta. \tag{6.6}$$

Because $|z||\theta| \geq |z\theta|$, it suffices to require $\varsigma^{-1}\theta^2 + \beta|\theta| > 2\varsigma^{-1}|z||\theta|$, which would imply (6.6) is true. This leads to

$$\varsigma^{-1}|\theta| + \beta > 2\varsigma^{-1}|z|.$$

Since the above inequality must be satisfied for all $\theta \neq 0$, it follows that $\theta^* = 0$ is the solution when $|z| \leq \varsigma\beta/2$.

Now consider $\theta^* > 0$, and hence when $z > \varsigma\beta/2$. Suppose specifically that $0 < \theta^* < 2\alpha\beta$ so that $\lambda^* = \beta - \frac{\theta^*}{2\alpha}$. Then θ^* must satisfy

$$z = \theta^* + \frac{\varsigma\lambda^*}{2} = \theta^* + \frac{\varsigma\beta}{2} - \frac{\varsigma\theta^*}{4\alpha}.$$

Solving for θ^* yields

$$\theta^* = \frac{4\alpha}{4\alpha - \varsigma}\left(z - \frac{\varsigma\beta}{2}\right). \tag{6.7}$$

Similarly for $-2\alpha\beta < \theta^* < 0$, and hence $z < -\varsigma\beta/2$, we obtain

$$\theta^* = \frac{4\alpha}{4\alpha - \varsigma}\left(z + \frac{\varsigma\beta}{2}\right). \tag{6.8}$$

Thus,

$$\theta^* = \begin{cases} 0 & \text{if } |z| < \frac{\varsigma\beta}{2} \\ \frac{4\alpha}{4\alpha-\varsigma}\left(z - sign(z)\frac{\varsigma\beta}{2}\right) & \text{if } \frac{\varsigma\beta}{2} \leq |z| < 2\alpha\beta \\ z & \text{if } |z| \geq 2\alpha\beta \end{cases}, \tag{6.9}$$

where the last case follows from the fact that if $|\theta^*| > 2\alpha\beta$, then $\lambda^* = 0$, in which case θ^* is the minimizer of $\varsigma^{-1}(\theta - z)^2 + \alpha\beta^2$ which occurs at $\theta^* = z$. □

Figure 6.1 illustrates the convexity of the (6.3) with $\varsigma = 1$, and thus a convexity "breakpoint" at $\alpha = 0.25$. The two rows and colors represent different values of β (0.25 and 1) while α is varied within rows (0.01, 0.25, 0.50); $\lambda \in (0, 10)$ and $\theta \in (-5, 5)$ are varied on the plot axes. In the first column, where $\alpha < 0.25$, the red line connecting two points on the surface lies underneath the surface, illustrating nonconvexity. The second column with $\alpha = 0.25$ shows the line connecting two points which appears to be tangent to the surface. The third column shows the line completely above the surface, illustrating convexity when $\alpha > 0.25$.

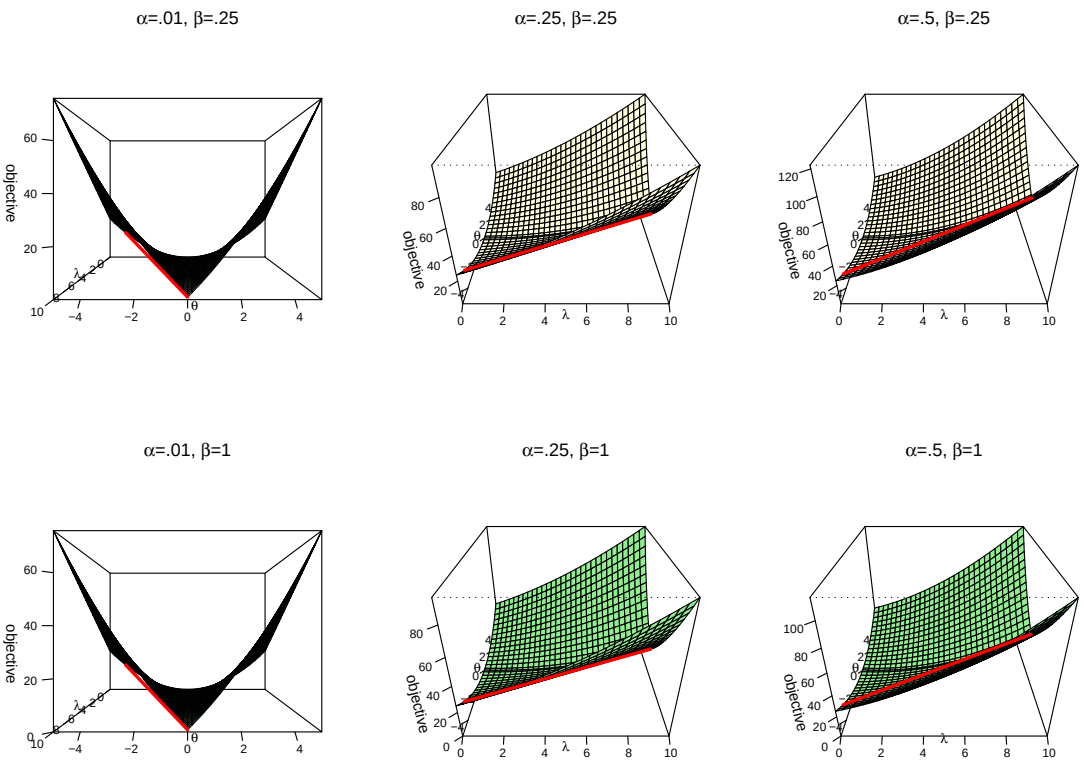

Figure 6.1: Convexity of HPLASSO-penalized objective function when $\alpha > 1/4$ for two different values of β (rows) and three different values of α (columns).

6.1.1 Connection with Minimax Concave Penalty

The right-hand-side of (6.4) acts as a thresholding function on z, say $\eta_{HP}(z;\alpha,\beta)$ and has the same form as the univariate MCP thresholding estimator of Zhang (2010) with $\varsigma = 2$, $2\alpha = a$ and $\beta = \lambda$. As shown above, the univariate HPLASSO thresholding estimator $\hat{\theta}_{HP}$ in (6.4) results from minimizing (6.3) simultaneously in θ and λ. However, the univariate MCP thresholding estimator,

$$\hat{\theta}_M \;:=\; \eta_M(z;a,\lambda) = \begin{cases} 0 & \text{if } |z| < \lambda \\ \frac{a}{a-1}\left(z - sign(z)\lambda\right) & \text{if } \lambda \le |z| < a\lambda \\ z & \text{if } |z| \ge a\lambda \end{cases}, \qquad (6.10)$$

results from minimizing the following objective function in θ only:

$$H(\theta;z) = \frac{1}{2}(\theta - z)^2 + \lambda\left(|\theta| - \frac{\theta^2}{2a\lambda}\right) I(|\theta| < a\lambda) + \frac{a\lambda^2}{2} I(|\theta| \ge a\lambda), \qquad (6.11)$$

for $\lambda > 0$ and $a > 1$. For the purpose of illustration, consider the following reparameterization of the univariate MCP objective function:

$$H(\theta; z) = \frac{1}{2}(\theta - z)^2 + \beta \left(|\theta| - \frac{\theta^2}{4\alpha\beta} \right) I(|\theta| < 2\alpha\beta) + \alpha\beta^2 I(|\theta| \geq 2\alpha\beta), \quad (6.12)$$

where $2\alpha = a$ and $\beta = \lambda$. Note that the solution for λ of (6.3) is $(\beta - \frac{|\theta|}{2\alpha})_+$. Setting $\lambda = (\beta - \frac{|\theta|}{2\alpha})_+$ and $\varsigma = 2$ in (6.3) yields an objective function of the form (6.12). Specifically, we have

$$\begin{aligned} G(\theta, \lambda; z) &= \frac{1}{2}(\theta - z)^2 + \left(\beta - \frac{|\theta|}{2\alpha} \right)_+ |\theta| + \alpha \left(\beta - (\beta - \frac{|\theta|}{2\alpha})_+ \right)^2 \\ &= \frac{1}{2}(\theta - z)^2 + \beta \left(|\theta| - \frac{\theta^2}{2\alpha\beta} + \frac{\theta^2}{4\alpha\beta} \right) I(|\theta| < 2\alpha\beta) \\ &\quad + \alpha\beta^2 I(|\theta| \geq 2\alpha\beta), \end{aligned}$$

which is equivalent to $H(\theta; z)$ in (6.12). Thus, $H(\theta; z)$ can be considered a profiled version of $G(\theta, \lambda; z)$ with $\lambda = \left(\beta - \frac{|\theta|}{2\alpha} \right)_+$.

Figure 6.2 shows the MCP-penalized objective function for various parameter values. Notice that the objective function is only convex in θ and not λ. This is in contrast to the HPLASSO-penalized objective function (see Figure 6.1), which is convex in both θ and λ for certain values of α. Indeed, the $\alpha(\beta - \lambda)^2$ term "convexifies" $\varsigma^{-1}(\theta - z)^2 + \lambda|\theta|$, considered jointly as a function of θ and λ, when $\alpha > \varsigma/4$.

6.2 Univariate Thresholding Estimators and Risk

In this section we compare the risks (assuming normality of z) of the HPLASSO/MCP univariate thresholding estimator (6.4) to three other well-known univariate estimators. Since the parameterization of the MCP penalty is related to the SCAD penalty, the MCP

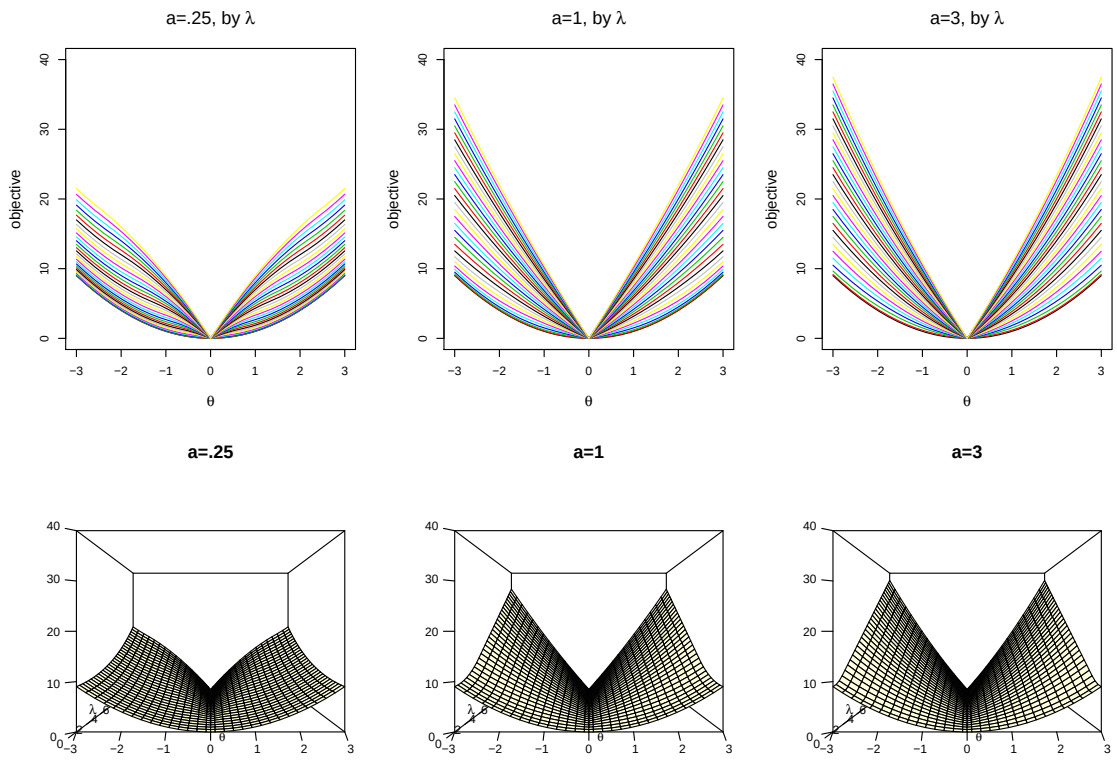

Figure 6.2: MCP-penalized objective function, notably not convex in λ for any value of a. Top: lines indicate objective functions for different values of λ at a specified a. Bottom: perspective plot versions of plots on top.

(a, λ) parameterization defined earlier is used below, with the estimator denoted here by $\hat{\theta}_M$ rather than by $\hat{\theta}_{HP}$.

In the univariate case, the univariate SCAD thresholding estimator is given by

$$\eta_{SC}(z, \lambda > 0, a > 2) = \begin{cases} sign(z)(|z| - \lambda)_+ & \text{if } |z| \leq 2\lambda \\ \frac{(a-1)z - sign(z)a\lambda}{a-2} & \text{if } 2\lambda < |z| \leq a\lambda \\ z & \text{if } |z| \geq a\lambda \end{cases}$$

(Fan and Li, 2001), whereas the SOFT and HARD thresholding estimators (written in capital letters henceforth to maintain consistency with capitalized acronyms SCAD and HPLASSO/MCP) are obtained by

$$\eta_S(z, \delta > 0) = sign(z)(|z| - \delta)_+,$$

$$\eta_H(z, \gamma > 0) = zI(|z| \geq \gamma).$$

As noted in Gao and Bruce (1997) and Zhang (2010), the MCP thresholding estimator approaches the SOFT thresholding estimator as $a \to \infty$; it approaches the HARD thresholding estimator as $a \to 1$ (see top row of Figure 6.3).

6.2.1 Theoretical Risk Formulae

Assuming $z = \theta + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, the risks for the HPLASSO/MCP, SCAD, SOFT, and HARD thresholding estimators are given below, and are plotted as a function of θ in Figure 6.3 (bottom). We take λ to be the 'zeroed-bandwidth' tuning parameter for all thresholding rules henceforth, and let $\phi(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\epsilon^2}{2\sigma^2}\right\}$ and $\Phi(\epsilon) = \int_{-\infty}^{\epsilon} \phi(u)du$. Note that the univariate risks for the HARD and SOFT thresholding estimators have been previously established (e.g., Droge, 1993, 1998; Donoho and Johnstone, 1994).

HPLASSO/MCP Thresholding Estimator: $\hat{\theta}_M = \eta_M(z; \lambda, a)$

$$
\begin{aligned}
r(\hat{\theta}_M, \theta) \;=\; & \sigma^2 + [\sigma^2 + \theta^2][\Phi(-a\lambda - \theta) - \Phi(a\lambda - \theta)] \\
& + \left(\frac{a}{a-1}\right)^2 \sigma^2[\Phi(a\lambda - \theta) - \Phi(-a\lambda - \theta) + \Phi(-\lambda - \theta) - \Phi(\lambda - \theta)] \\
& + \left(\frac{a}{a-1}\right)^2 (\theta + \lambda)^2[\Phi(-\lambda - \theta) - \Phi(-a\lambda - \theta)] \\
& + \left(\frac{a}{a-1}\right)^2 (\theta - \lambda)^2[\Phi(a\lambda - \theta) - \Phi(\lambda - \theta)] \\
& + \frac{2\theta(a\lambda + \theta)}{a-1}\Phi(-a\lambda - \theta) + \frac{2\theta(a\lambda - \theta)}{a-1}\Phi(a\lambda - \theta) \\
& - \frac{2\theta a(\lambda + \theta)}{a-1}\Phi(-\lambda - \theta) - \frac{2\theta a(\lambda - \theta)}{a-1}\Phi(\lambda - \theta) \\
& + \sigma^2[(a\lambda + \theta)\phi(a\lambda + \theta) + (a\lambda - \theta)\phi(a\lambda - \theta)] \\
& - \frac{\sigma^2 a(a-2)(a\lambda + \theta)}{(a-1)^2}\phi(a\lambda + \theta) - \frac{\sigma^2 a(a-2)(a\lambda - \theta)}{(a-1)^2}\phi(a\lambda - \theta) \\
& + \frac{\sigma^2 a(a\theta - a\lambda - 2\theta)}{(a-1)^2}\phi(\lambda + \theta) - \frac{\sigma^2 a(a\theta + a\lambda - 2\theta)}{(a-1)^2}\phi(\lambda - \theta)
\end{aligned}
$$

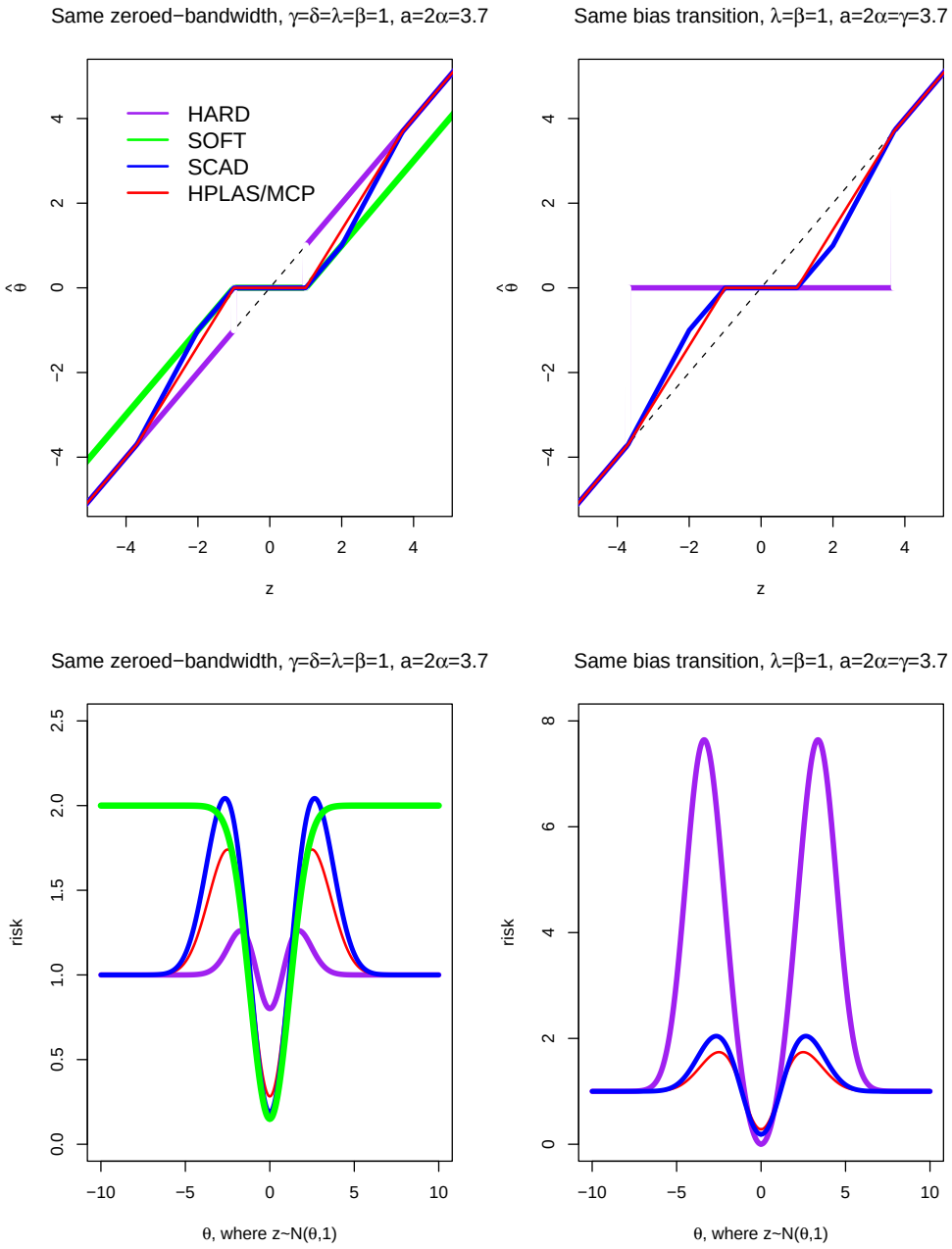

Figure 6.3: Correspondence between parameters in terms of the length of the thresholded region (zeroed-bandwidth) and the point at which the bias disappears (bias transition), as well as their univariate risks assuming $z \sim N(\theta, \sigma^2)$. Note: the SCAD and HPLASSO/MCP thresholding functions in the top row have the same zeroed bandwidth and same bias transition in both plots; it is the HARD thresholding rule that changes according to the figure headings.

SCAD Thresholding Estimator: $\hat{\theta}_{SC} = \eta_{SC}(z;\lambda,a)$

$$
\begin{aligned}
r(\hat{\theta}_{SC},\theta) &= \sigma^2 + \sigma^2[\Phi(-a\lambda-\theta) - \Phi(a\lambda-\theta)] - \sigma^2\theta[\phi(2\lambda+\theta) - \phi(2\lambda-\theta)] \\
&\quad +[\sigma^2+\lambda^2-\theta^2][\Phi(-\lambda-\theta) - \Phi(\lambda-\theta)] \\
&\quad +(\sigma^2+\lambda^2)[\Phi(2\lambda-\theta) - \Phi(-2\lambda-\theta)] \\
&\quad + \left(\frac{a-1}{a-2}\right)^2 \sigma^2[\Phi(a\lambda-\theta) - \Phi(-a\lambda-\theta)] \\
&\quad + \left(\frac{a-1}{a-2}\right)^2 \sigma^2[\Phi(-2\lambda-\theta) - \Phi(2\lambda-\theta)] \\
&\quad + \left(\frac{a\lambda-\theta}{a-2}\right)^2 [\Phi(a\lambda-\theta) - \Phi(2\lambda-\theta)] \\
&\quad + \left(\frac{a\lambda+\theta}{a-2}\right)^2 [\Phi(-2\lambda-\theta) - \Phi(-a\lambda-\theta)] \\
&\quad +\sigma^2[(a\lambda+\theta)\phi(a\lambda+\theta) + (a\lambda-\theta)\phi(a\lambda-\theta)] \\
&\quad -\sigma^2[(\lambda-\theta)\phi(\lambda+\theta) + (\lambda+\theta)\phi(\lambda-\theta)] \\
&\quad -\frac{\sigma^2(a-1)(a-3)(a\lambda+\theta)}{(a-2)^2}\phi(a\lambda+\theta) \\
&\quad -\frac{\sigma^2(a-1)(a-3)(a\lambda-\theta)}{(a-2)^2}\phi(a\lambda-\theta) \\
&\quad +\frac{\sigma^2(a-1)(a\theta-2\lambda-3\theta)}{(a-2)^2}\phi(2\lambda+\theta) \\
&\quad -\frac{\sigma^2(a-1)(a\theta+2\lambda-3\theta)}{(a-2)^2}\phi(2\lambda-\theta)
\end{aligned}
$$

SOFT Thresholding Estimator: $\hat{\theta}_S = \eta_S(z;\lambda)$

$$
\begin{aligned}
r(\hat{\theta}_S,\theta) &= \sigma^2 + \lambda^2 + [\theta^2-\sigma^2-\lambda^2][\Phi(\lambda-\theta) - \Phi(-\lambda-\theta)] \\
&\quad -\sigma^2[(\lambda-\theta)\phi(\lambda+\theta) + (\lambda+\theta)\phi(\lambda-\theta)]
\end{aligned}
$$

HARD Thresholding Estimator: $\hat{\theta}_H = \eta_H(z;\lambda)$

$$
\begin{aligned}
r(\hat{\theta}_H,\theta) &= \sigma^2 + [\theta^2-\sigma^2][\Phi(\lambda-\theta) - \Phi(-\lambda-\theta)] \\
&\quad +\sigma^2[(\lambda+\theta)\phi(\lambda+\theta) + (\lambda-\theta)\phi(\lambda-\theta)]
\end{aligned}
$$

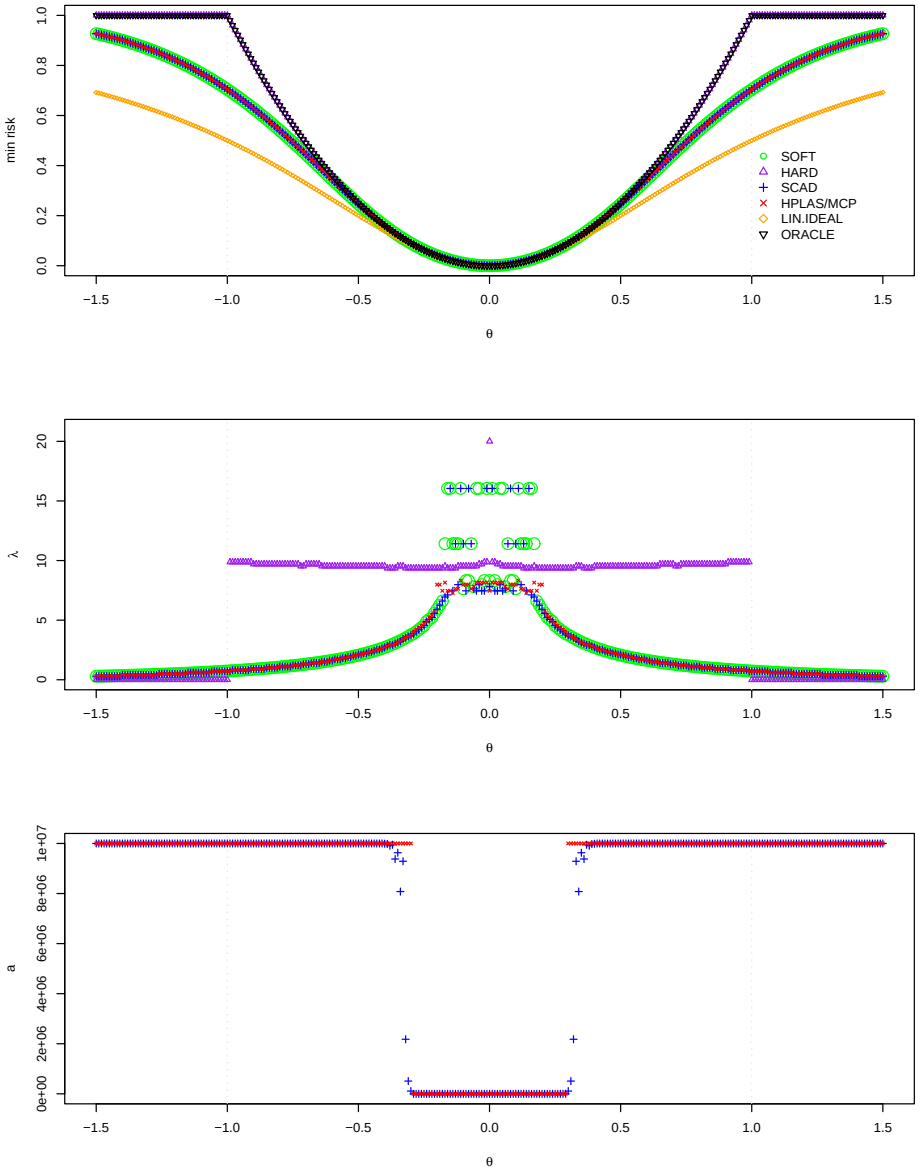

Figure 6.4: Minimum univariate risks (in a and/or λ) assuming $z \sim N(\theta, 1)$.

6.2.2 Minimum Risks as a Function of Tuning Parameters

Minimizing over a grid of possible tuning parameters (a and/or λ), the univariate minimum risks for the four estimators are displayed in Figure 6.4. For large values of θ, the risk is minimized by taking $\lambda \to 0$ for all four thresholding estimators.

Interestingly, minimizing the univariate risk for SCAD and MCP in a and λ reduces to the minimum univariate risk for the SOFT thresholding estimator. For larger values of $\theta,$ the minimum risk for SCAD and MCP occurs when $a \to \infty$ and λ is (essentially) the same as for the minimum SOFT risk. While the tuning parameters are not necessarily identical across SOFT, SCAD, and MCP, the minimum risks are seemingly identical. For smaller values of $\theta,$ the optimal a parameters are selected to be the minimum considered value ($a = 2.01$ for SCAD and $a = 1.01$ for MCP). From the perspective of (univariate) risk minimization, these results suggest that with optimal selection of tuning parameters, the SCAD or MCP thresholding estimators are just as good as the SOFT thresholding estimator. Of course, the minimum risk perspective does not tell the whole story, as selection accuracy (in terms of correctly identifying the zero and nonzero coefficients) is also a very important issue within the thresholding community.

As a benchmark for comparison, we also provide the univariate oracle risk and ideal linear risk as defined in Wasserman (2006). The oracle risk is for an estimator that knows whether the risk is minimized for nonzero or zero values, and is given by $\min(\sigma^2, \theta^2)$ (e.g., Wasserman, 2006, pg 172). This is identical to the univariate minimum risk for the HARD thresholding estimator. The ideal linear risk is for estimators of the form bZ is given by $\theta^2/(1 + \theta^2)$ (e.g., Wasserman, 2006, pg 155). The ideal linear risk is much lower than any of the other risks considered in the univariate case, although only at the extremes. In fact, all optimally tuned risks generally behave quite similarly under 'near sparsity' on the range [-0.5,0.5]; the differences become more prominent as the magnitude of θ increases.

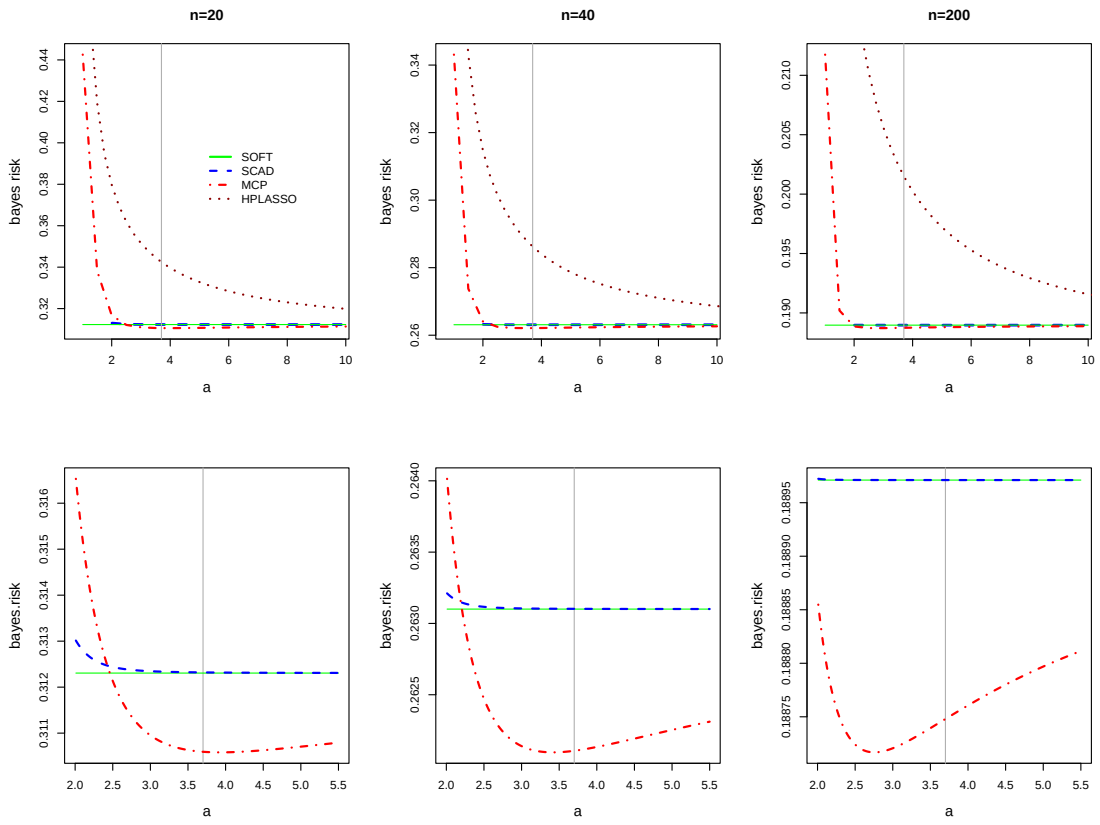

Figure 6.5: Bayes risk plotted against parameter a for three values of $\lambda = \sqrt{2\log n}$ ($\beta = \sqrt{2\log n}$ for HPLASSO); bottom row plots are zoomed-in versions of top row plots on the interval [2,5.5].

6.2.3 Bayes Risk as a Function of a

Using the univariate risks from the previous section, the Bayes risk, $r(\pi, \hat{\theta}) = \int r(\hat{\theta}, \theta)\pi(\theta)d\theta$, was approximated using Monte Carlo integration for the SOFT, SCAD, and HPLASSO/MCP thresholding estimators. As indicated by the notation, the Bayes risk depends on the marginal prior distribution for θ, i.e., $\pi(\theta)$. The risks for specific choices of prior distributions (dependent upon specific choices of λ or β), plotted as a function of $a = 2\alpha$, are shown in Figure 6.5.

The $DoubExp(\lambda)$ prior distribution for θ is a natural choice for the SOFT thresholding estimator, as the latter is the MAP estimator for θ under this prior for objective

$\frac{1}{2}(\theta - z)^2 + \lambda|\theta|$. Since the SCAD and MCP estimators approach the SOFT thresholding estimator in the limit as $a \to \infty$, we considered the $DoubExp(\lambda)$ for the SCAD and MCP thresholding estimators, as well. The HARD thresholding estimator was not considered, as the appropriate choice of prior was less obvious. In Figure 6.5, we considered three choices of $\lambda = \sqrt{2 \log n}$ for $n \in \{20, 40, 200\}$; note that the risks were computed for a univariate observation and the value of n was used only to define λ. As the SOFT thresholding estimator does not depend on a, the same Bayes risk value was plotted along the a axis for these cases in Figure 6.5. The Bayes risk for SCAD quickly approaches this value as a increases, whereas the MCP Bayes risk evidently drops below this value and reaches its minimum somewhere between 2.5 and 4, depending on λ.

Since the MCP thresholding estimator can be motivated from a hierarchical structure, we also considered the Bayes risk under the (proper) joint prior defined by

$$\pi(\theta|\lambda, \alpha, \beta) \sim \lambda \exp\{-\lambda|\theta|\}, \;\; \pi(\lambda|\alpha, \beta) \propto \exp\{-\alpha(\lambda - \beta)^2\}. \tag{6.13}$$

Note that $\pi(\lambda|\alpha, \beta)$ in (6.13) is proportional to a folded normal distribution with mean β and variance $(2\alpha)^{-1}$ ($= a^{-1}$ in the MCP parameterization), and does not include the extra λ factor as in (6.1). The resulting marginal prior is very much related to the quasi-Cauchy distribution as described by Johnstone and Silverman (2004, 2005), and subsequently used in Schifano (2007), Zhang et al. (2010a), and Griffin and Brown (2005, 2007). In fact, for $\beta = 0$, the marginal prior is exactly the quasi-Cauchy distribution. The Bayes risk under choice of prior (6.13) with $\beta = \sqrt{2 \log n}$, $n \in \{20, 40, 200\}$, plotted as a function of $a = 2\alpha$, is labeled HPLASSO in Figure 6.5. Not surprisingly, since $1/a$ is related to the variance of λ, low values of a result in high Bayes risk. Eventually, as a grows so that the variance for λ is essentially zero, the Bayes risk for HPLASSO reaches the same Bayes risk levels as SOFT, SCAD, and MCP.

The $DoubExp(\lambda)$ prior for θ results in small Bayes risk for SCAD and MCP at $a =$

3.7, the recommended value in Fan and Li (2001). They consider a different, $N(0, a\lambda)$ prior, and show that the minimum risk for SCAD occurs at 3.7; we unfortunately have not been able to reproduce this result. Interestingly, though, the use of a hierarchically specified prior suggests that a higher value of a (α) is required.

6.3 Selection of Tuning Parameters

In practice, the true means $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$ are unknown, yet the problem of selecting tuning parameters still remains. Minimization of Stein's Unbiased Risk Estimators (SURE) and Generalized Cross-Validation (GCV) error are common ways for selecting tuning parameters, collected in say $\boldsymbol{\lambda}$, that are based solely on the observed data $\mathbf{z} = (z_1, \ldots, z_n)'$ and estimates $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}} = (\eta(z_1; \boldsymbol{\lambda}), \ldots, \eta(z_n; \boldsymbol{\lambda}))'$ for some generic thresholding rule η.

For a vector of independent observations $\mathbf{z}$ where $z_i \sim N(\theta_i, \sigma^2)$, $i = 1, \ldots, n$, the Stein's Unbiased Risk Estimator (SURE) of Stein (1981), as the name implies, provides an unbiased estimator of risk under certain weak differentiability conditions. In our context, the SURE, a function of $\mathbf{z}$ only, unbiasedly estimates $\sum_{i=1}^{n} r(\hat{\theta}_i, \theta_i)$. In particular, the estimates take the form $n\sigma^2 + 2\sigma^2 div\ g(\mathbf{z}) + \|g(\mathbf{z})\|^2$ where $g(\mathbf{z}) = \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}} - \mathbf{z}$. The SURE criteria for the HPLASSO/MCP, SCAD, and SOFT thresholding estimators are given below. Note that the SOFT SURE has previously appeared in the literature (e.g., Wasserman, 2006, pg 152). The HPLASSO SURE can be seen as a special case of the SURE formula provided in Strawderman and Wells (2010), and also coincides with the SURE formula given in Zhang (2010, Section 5) for $\mathbf{X}$, n and p in their notation set equal to $\mathbf{I}$, one, and n, respectively. The SCAD SURE can similarly be obtained using the results in Section 5 of Zhang (2010). Figure 6.6 demonstrates that the estimators

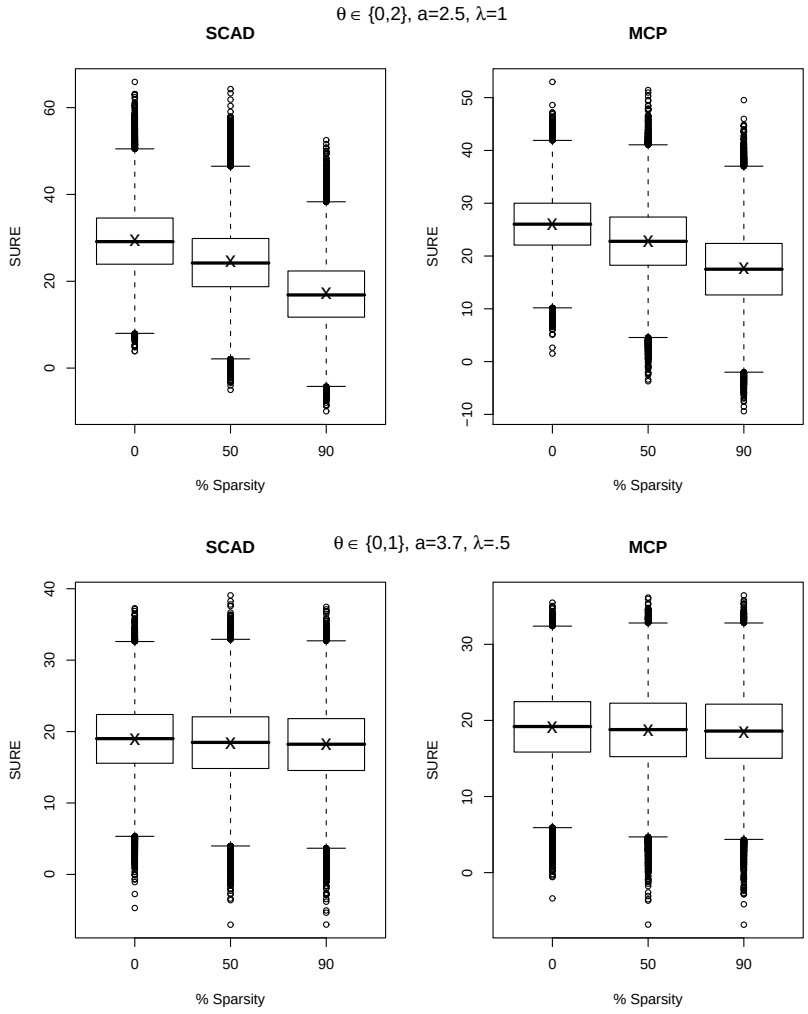

Figure 6.6: For fixed θ, a, λ, SURE is unbiased, with true risk indicated by "x" in each boxplot.

for SCAD and MCP are indeed unbiased. Also note that the HARD SURE can not be computed, as the estimator is not weakly differentiable.

HPLASSO/MCP SURE:

$$\begin{aligned} R_M(\mathbf{z}, \lambda, a) &= \sum_{i=1}^{n} \Big(\sigma^2 + (|z_i|^2 - 2\sigma^2) I(|z_i| \leq \lambda) \\ &\quad + \left[\frac{2\sigma^2}{a-1} + \frac{(|z_i| - a\lambda)^2}{(a-1)^2} \right] I(\lambda < |z_i| \leq a\lambda) \Big) \end{aligned}$$

SCAD SURE:

$$R_{SC}(\mathbf{z},\lambda,a) = \sum_{i=1}^{n}\Big(\sigma^2+(|z_i|^2-2\sigma^2)I(|z_i|\le\lambda)+\lambda^2 I(\lambda<|z_i|\le 2\lambda) + \left[\frac{2\sigma^2}{a-2}+\frac{(|z_i|-a\lambda)^2}{(a-2)^2}\right]I(2\lambda<|z_i|\le a\lambda)\Big)$$

SOFT SURE:

$$R_S(\mathbf{z},\lambda) = \sum_{i=1}^{n}\big(\sigma^2-2\sigma^2 I(|z_i|\le\lambda)+\min(z_i^2,\lambda^2)\big)$$

As described in Li (1985), the generalized cross validation (GCV) criterion in our context is given by

$$GCV_n(\boldsymbol{\lambda}) = \frac{n^{-1}\|\mathbf{z}-\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}\|^2}{\big[1-n^{-1}\text{tr}(\mathbf{M}_n(\boldsymbol{\lambda}))\big]^2}, \tag{6.14}$$

where $\mathbf{M}_n$ is an $n\times n$ matrix associated with $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$ such that $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}=\mathbf{M}_n(\boldsymbol{\lambda})\mathbf{z}$. Conveniently, the HPLASSO/MCP, SCAD, and SOFT thresholding estimators can each be written in the form $\omega(|z|,\boldsymbol{\lambda})z$ so that $\mathbf{M}_n = diag(\omega(|z_i|,\boldsymbol{\lambda}), i=1,\ldots,n)$. Specifically,

$$\omega_S(x,\lambda) = \begin{cases} 0 & x\le\lambda \\ 1-\frac{\lambda}{x} & x>\lambda \end{cases} \tag{6.15}$$

$$\omega_{SC}(x,a,\lambda) = \begin{cases} 0 & x\le\lambda \\ 1-\frac{\lambda}{x} & \lambda<x\le 2\lambda \\ \frac{a-1}{a-2}-\frac{a\lambda}{(a-2)x} & 2\lambda<x\le a\lambda \\ 1 & x>a\lambda \end{cases}$$

$$\omega_M(x,a,\lambda) = \begin{cases} 0 & x\le\lambda \\ \frac{a}{a-1}\left(1-\frac{\lambda}{x}\right) & \lambda<x\le a\lambda \\ 1 & x>a\lambda \end{cases}$$

for the HPLASSO/MCP, SCAD, and SOFT thresholding estimators, respectively.

Interestingly, Li (1985) demonstrates how one can derive GCV from SURE for Stein estimates associated with estimators $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}} = \mathbf{M}_n(\boldsymbol{\lambda})\mathbf{z}$, where $\mathbf{M}_n$ is symmetric. He defines the Stein estimate associated with $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$ as

$$\hat{\boldsymbol{\theta}}^S_{\boldsymbol{\lambda}} = \mathbf{z} - \frac{\sigma^2}{\mathbf{z}'\mathbf{B}_n(\boldsymbol{\lambda})\mathbf{z}}\mathbf{A}_n(\boldsymbol{\lambda})\mathbf{z}, \tag{6.16}$$

where $\mathbf{A}_n(\boldsymbol{\lambda}) = \mathbf{I}_n - \mathbf{M}_n(\boldsymbol{\lambda})$ and $\mathbf{B}_n(\boldsymbol{\lambda}) = [(\text{tr}\mathbf{A}_n(\boldsymbol{\lambda})) \cdot \mathbf{I}_n - 2\mathbf{A}_n(\boldsymbol{\lambda})]^{-1}\mathbf{A}_n(\boldsymbol{\lambda})^2$. Under assumptions on the eigenvalues of $\mathbf{A}_n(\boldsymbol{\lambda})$ and using an approximation of $\mathbf{B}_n(\boldsymbol{\lambda})$, the SURE for estimator $\hat{\boldsymbol{\theta}}^S_{\boldsymbol{\lambda}}$ reduces to the GCV of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$ as in (6.14). Using the results of Li and Hwang (1984), similar connections between SURE and GCV can be obtained for estimators with non-symmetric $\mathbf{M}_n$. Li (1985) remarks that when the estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$ is 'good' in the sense that $\boldsymbol{\lambda}$ is appropriate, then the shrinkage factor $\frac{\sigma^2}{\mathbf{z}'\mathbf{B}_n(\boldsymbol{\lambda})\mathbf{z}}$ should be close one and $\hat{\boldsymbol{\theta}}^S_{\boldsymbol{\lambda}} \approx \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$. In this case, their respective SURE formulas should approximately coincide, and GCV can also be used to approximate the SURE of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$. This observation suggests the strong possibility of similar tuning parameter selection performance from GCV and SURE criteria minimization.

Thus, we undergo a simulation study designed to address tuning parameter selection for SOFT, SCAD, and HPLASSO/MCP thresholding estimators using the SURE and GCV criteria to select one or both tuning parameters λ and a. We evaluate performance of the different selection criteria using the average empirical risk (AER) as our metric. We define AER as follows:

$$AER(\boldsymbol{\theta}) = B^{-1}\sum_{b=1}^{B} err^{(b)}(\boldsymbol{\theta}) \tag{6.17}$$

where $err^{(b)}(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(b)}_{\boldsymbol{\lambda}}\|^2$ is the empirical risk for estimator $\hat{\boldsymbol{\theta}}^{(b)}_{\boldsymbol{\lambda}}$ for the b^{th} dataset.

6.3.1 Data Generation and Set-up

Consider each $\theta_i \sim (1-\pi)N(\mu,1) + \pi\delta_0(\mu)$, $i = 1,\ldots,n$ where the first component of the mixture is a normal distribution with mean μ and the second component is a point mass at zero to induce sparsity in $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)'$. For a given μ in the range [-5,5] and $\pi \in \{0, 0.5, 0.9\}$, a single set of θ_i, $i = 1,\ldots,n$ is randomly generated according to the mixture prior for $n = 20$. From this single set of $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)'$ associated with the given μ and π, we generate $B = 500$ vectors indexed by b, $\mathbf{z}^{(b)}$, such that $z_i^{(b)} \sim N(\theta_i, 1)$, $i = 1,\ldots,n = 20$.

For SCAD and HPLASSO/MCP, $\boldsymbol{\lambda} = (\lambda, a)$ so two parameters need to be selected. We consider the following strategies of selection for a given dataset:

(i.) Minimize SURE in both λ and a,

(ii.) Select data adaptive λ, minimize SURE in a for selected λ,

(iii.) Select data adaptive λ, minimize GCV in a for selected λ,

(iv.) Fix $a = 3.7$, minimize SURE in λ,

(v.) Fix $a = 3.7$, select data adaptive λ.

Several comments are in order regarding these strategies. First, when selecting λ for each replicate dataset $\mathbf{z}^{(b)}$ in strategies (ii.), (iii.), and (v.) we use the estimation method of Johnstone and Silverman (2004). They provide an empirical Bayes, and thus data-adaptive, estimate for λ using independent mixture priors

$$\pi(\theta_i) = (1-w)\delta_0(\theta_i) + w\gamma(\theta_i) \tag{6.18}$$

with γ representing either the Laplace or quasi-Cauchy distribution. We use the quasi-Cauchy distribution as the non-point mass component (i.e., $\gamma(\cdot)$) to estimate λ, subject

to the constraint that $\lambda \leq \sqrt{2 \log n}$. Despite the implicit use of the posterior median in its calculation, Johnstone and Silverman (2004, page 1608) remark that the resulting estimator for λ (computed using the `tfromx` function within the *EbayesThresh* R package) retains its desirable data adaptivity properties for other thresholding estimators (e.g., SOFT, HARD).

Second, we attempted two additional strategies (not listed above) which involved minimizing GCV_n in both λ and a, and minimizing GCV_n in λ for fixed $a = 3.7$. However, we quickly realized these strategies are not advisable for estimators such as SCAD and MCP in which no thresholding occurs, i.e., $\hat{\theta}_{\boldsymbol{\lambda},i} = z_i$ for all $i = 1, \ldots, n$. Upon examining (6.14), we immediately see that in such cases both the numerator and denominator equal zero. An obvious 'fix' would be to set 0/0=0, which would of course yield the minimum GCV_n. This, however, has the undesirable effect that we will always select parameters λ (and a, if applicable) such that $\hat{\theta}_{\boldsymbol{\lambda},i} = z_i$ for all $i = 1, \ldots, n$, despite the level of sparsity. The alternative of adding a fixed ϵ to both the numerator and the denominator also has its drawbacks, as the choice of ϵ and its placement indeed affect the selection process and resulting empirical risks. Thus, these strategies involving of selection of λ by minimizing GCV_n were not included. We remark, however, that we do set $GCV_n(\boldsymbol{\lambda}) = 0$ in strategy (iii.) in situations when no thresholding occurs. This is acceptable here, as we are selecting λ externally using the Johnstone and Silverman (2004) method.

Finally, for the fixed value of a, we chose $a = 3.7$ as it is both the suggested value of Fan and Li (2001), and has favorable Bayes risk as previously illustrated in Figure 6.5.

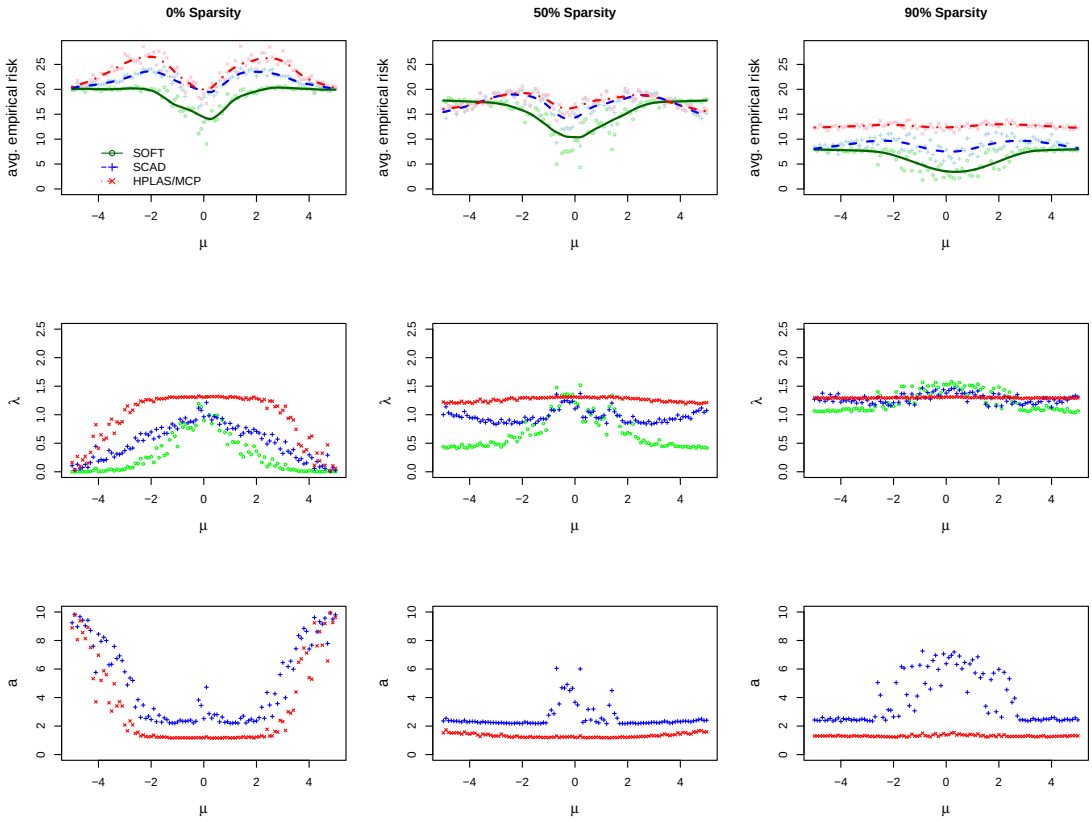

Figure 6.7: Average empirical risks for tuning parameters selected by minimizing SURE criteria (in λ and a; strategy (i.)) assuming $z_i \sim N(\theta_i, 1)$, $\theta_i \sim (1-\pi)N(\mu, 1)+\pi\delta_0(\mu)$, $i = 1, \ldots, n = 20$, at different levels of sparsity ($\pi \in \{0, .5, .9\}$). The second and third rows plot the average minimum λ and average maximum a for which the minimum SURE is achieved.

6.3.2 Results

The results for strategy (i.) are plotted in Figures 6.7 and 6.8. In each, the first row displays the AERs for the three estimators with their respective lowess-smoothed trends at the three levels of sparsity. The corresponding SURE-minimizing average λ and a parameters are displayed in the second and third rows, where the average is computed for each μ over the B datasets. However, it should be noted that the minimum unbiased estimator of risk for SCAD and HPLASSO/MCP can be achieved at multiple combinations of a and λ. Additionally, the minimum SURE can be achieved at a range of λ for

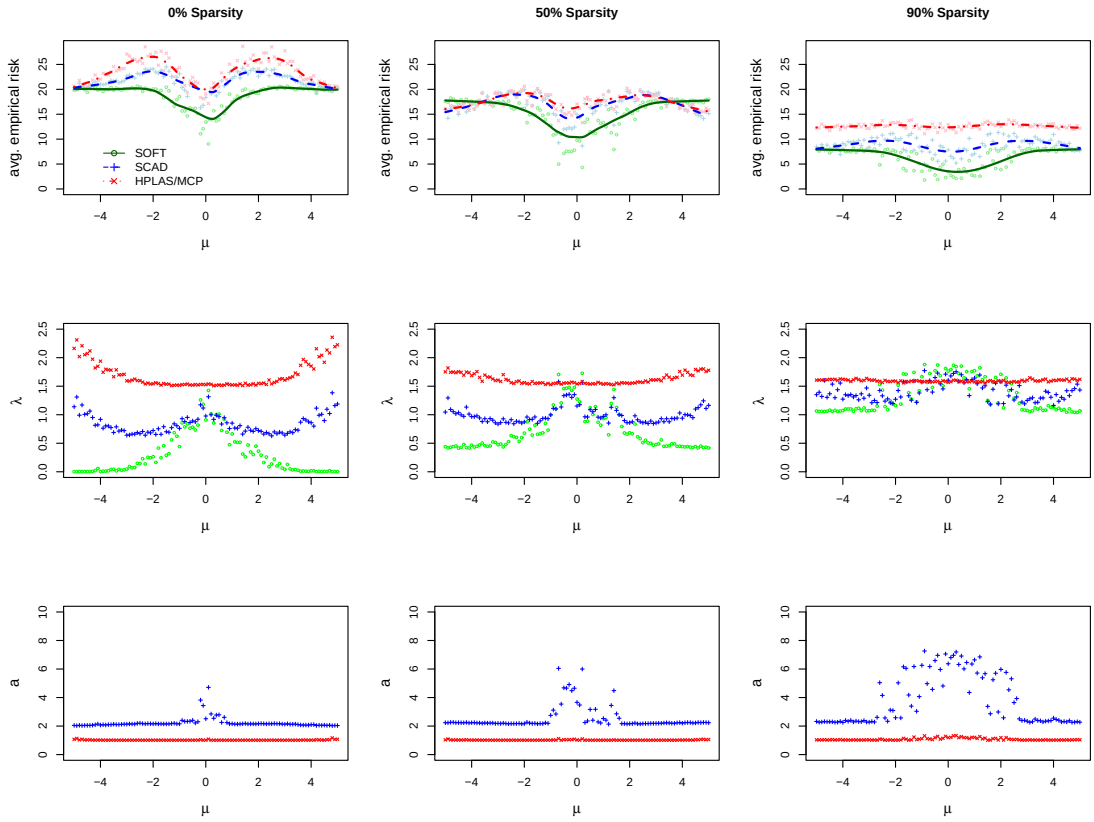

Figure 6.8: Average empirical risks for tuning parameters selected by minimizing SURE criteria (in λ and a; strategy (i.)) assuming $z_i \sim N(\theta_i, 1)$, $\theta_i \sim (1-\pi)N(\mu, 1)+\pi\delta_0(\mu)$, $i = 1, \ldots, n = 20$, at different levels of sparsity ($\pi \in \{0, .5, .9\}$). The second and third rows plot the average maximum λ and average maximum a for which the minimum SURE is achieved.

the SOFT thresholding estimator. Thus, the second and third rows of Figure 6.7 plot the average *minimum* λ and average *maximum* a for which the minimum estimate of risk is achieved; the second and third rows of Figure 6.8 plot the average *maximum* λ and average *maximum* a for which the minimum estimate of risk is achieved. Note that the maximum allowable value for a was set at ten while the maximum allowable value for λ was set to $\sqrt{2 \log n}$.

The analogous results for strategy (ii.) are plotted in Figure 6.9, where we plot the empirical Bayes-selected λ (which is the same for all thresholding rules for a given dataset) in the second row, and the average *maximum* a for which the minimum estimate

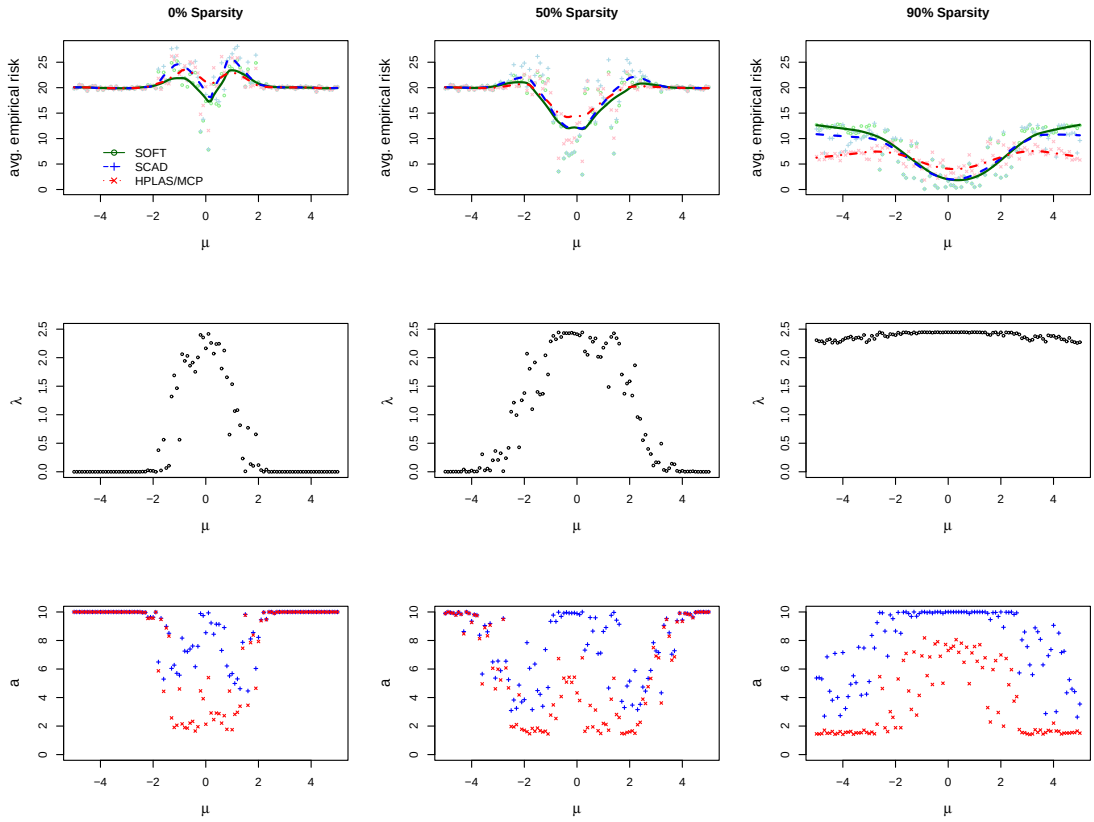

Figure 6.9: Average empirical risks for tuning parameter a selected by minimizing SURE criteria for data-adaptive λ (strategy (ii.)) assuming $z_i \sim N(\theta_i, 1)$, $\theta_i \sim (1-\pi)N(\mu, 1) + \pi\delta_0(\mu)$, $i = 1, \ldots, n = 20$, at different levels of sparsity ($\pi \in \{0, .5, .9\}$). The second and third rows plot the average empirical Bayes-selected λ and average maximum a for which the minimum SURE is achieved.

of risk is achieved. Interestingly, the shape of the smoothed-curves for the 0% sparsity case shares a similar form to the curves featured in the univariate risk plots (bottom of Figure 6.3). Note that since the SOFT thresholding estimator does not depend on a, no minimization was performed, but the AERs were still computed and plotted for the selected λ.

The empirical risks and tuning parameters for the GCV selection method (strategy (iii.)) are provided in Figure 6.10. These are strikingly similar to the empirical risks and tuning parameters selected by SURE in Figure 6.9. The similarity is further illustrated in Figure 6.11 (top row), which shows the average empirical risk differences.

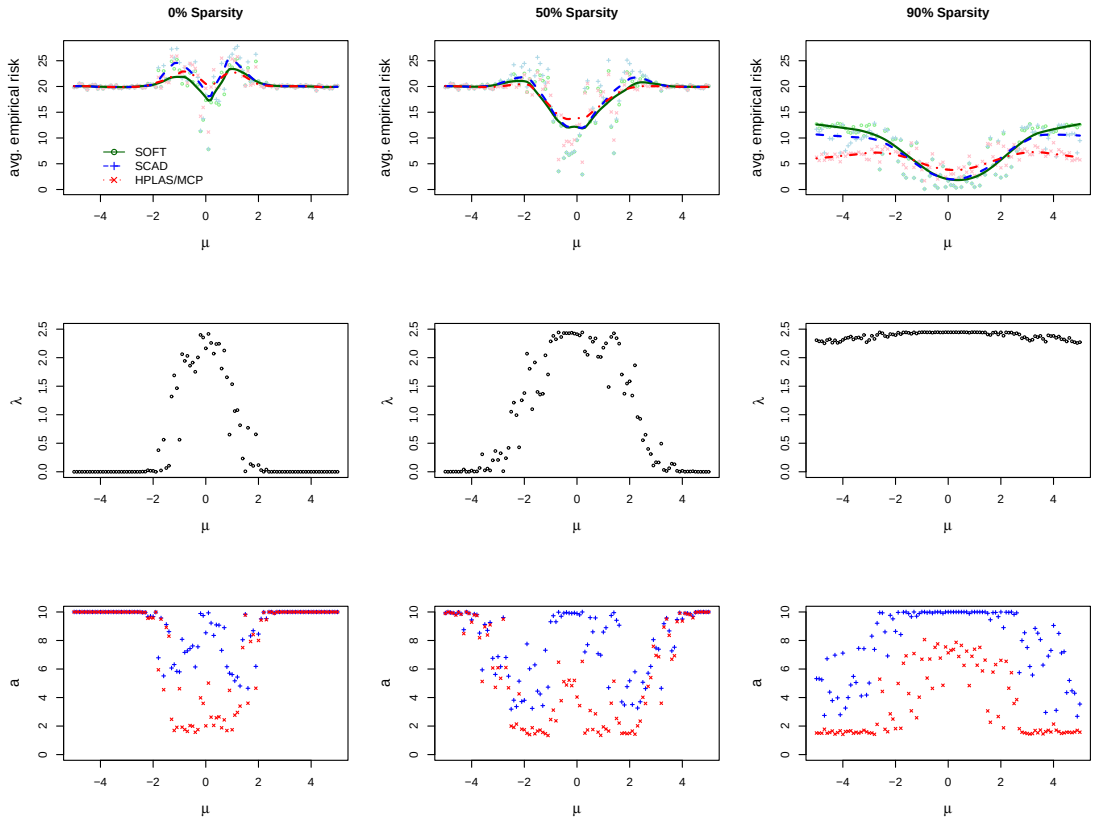

Figure 6.10: Average empirical risks for tuning parameter a selected by minimizing GCV criteria for data-adaptive λ (strategy (iii.)) assuming $z_i \sim N(\theta_i, 1)$, $\theta_i \sim (1 - \pi)N(\mu, 1) + \pi\delta_0(\mu)$, $i = 1, \ldots, n = 20$, at different levels of sparsity ($\pi \in \{0, .5, .9\}$). The second and third rows plot the average empirical Bayes-selected λ and average maximum a for which the minimum GCV is achieved.

The difference is 'AER (ii.) minus AER (iii.)', so smoothed curves that are higher than zero indicate GCV-based tuning parameters provide lower AERs than the SURE-based tuning parameters. The vertical scale is quite small, so the GCV-based parameters and SURE-based parameters are essentially equivalent. The differences are most noticeable, however, for the HPLASSO/MCP thresholding estimator at the high level of sparsity.

For comparison, the average empirical risk differences were also computed for the first and second strategies involving SURE; these differences are displayed in the bottom row of Figure 6.11. The difference is 'AER (ii.) minus AER (i)', so smoothed curves that are lower than zero indicate that the Johnstone and Silverman (2004) method of

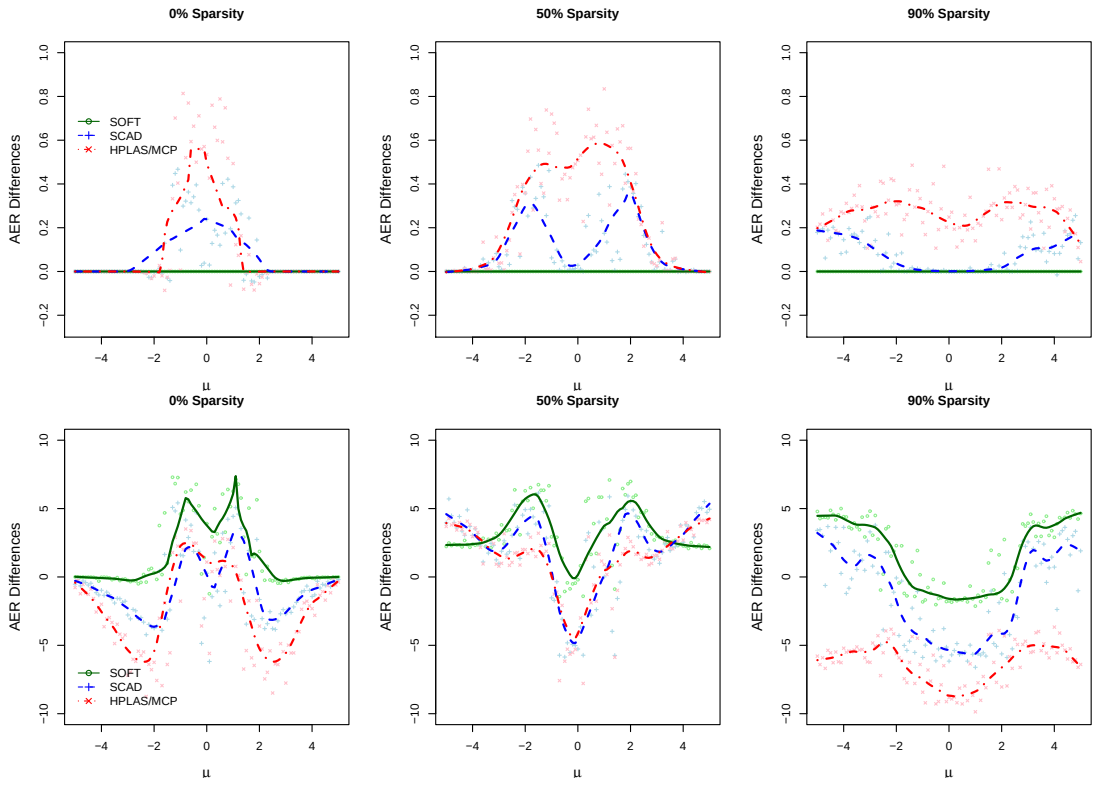

Figure 6.11: Differences in Average Empirical Risks (AER). Top: Difference is 'AER (ii.) minus AER (iii.)', so smoothed curves higher than zero indicate GCV-based tuning parameters have lower AER than SURE-based tuning parameters. Bottom: Difference is 'AER (ii.) minus AER (i.)', so smoothed curves higher than zero indicate the solely SURE-based tuning parameters have lower AER than the SURE-based a and empirical Bayes-selected λ.

selecting λ resulted in lower AERs than the SURE minimization in both λ and a. Here, neither strategy (i.) nor (ii.) really dominates the other in terms of AER, except in the 90% sparsity case, where the MCP smoothed AER curve lies completely below zero. This suggests that the external, data-adaptive selection of λ in conjunction with minimization of SURE in a yields better estimates than from minimizing SURE in λ and a simultaneously, at least for HPLASSO/MCP under high sparsity.

Figures 6.12 and 6.13 correspond to the fixed a strategies (iv.) and (v.), respectively. The structure of these plots is slightly different than the rest, as the plots for a are no longer needed. The second and third rows of Figure 6.12 both display the average λ

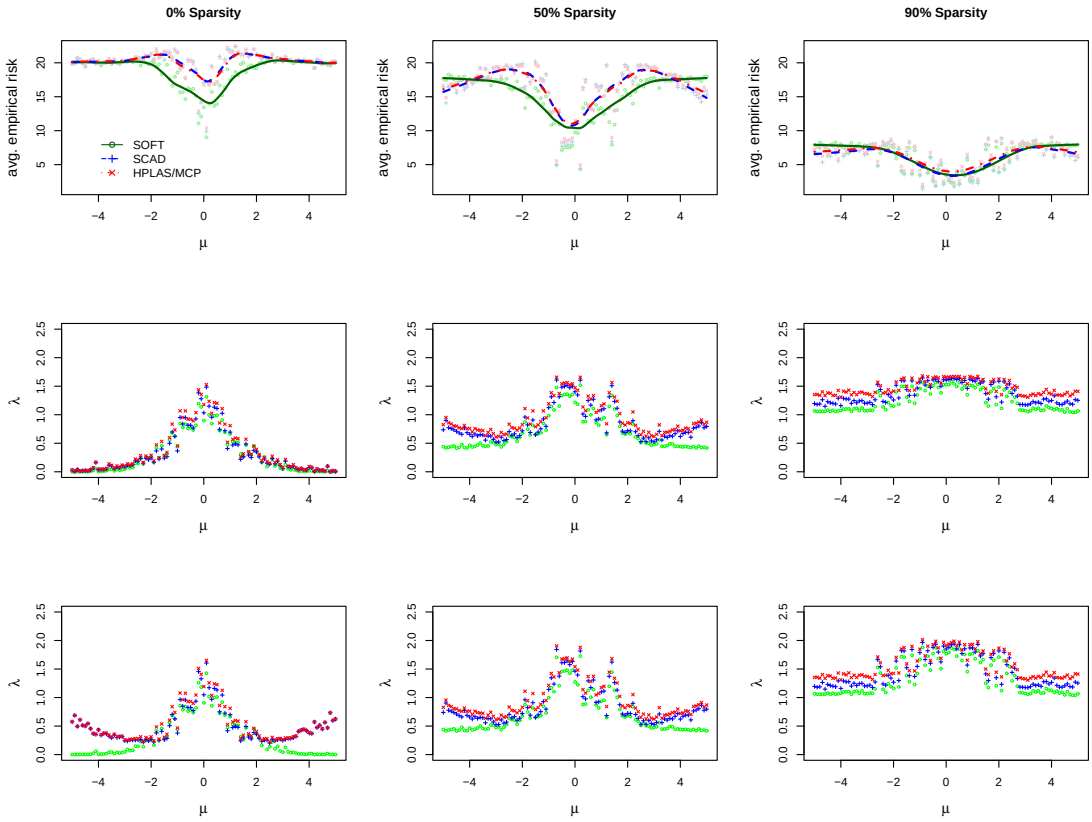

Figure 6.12: Average empirical risks for tuning parameter λ selected by minimizing SURE criteria for fixed $a = 3.7$ (strategy (iv.)), assuming $z_i \sim N(\theta_i, 1)$, $\theta_i \sim (1 - \pi)N(\mu, 1) + \pi\delta_0(\mu)$, $i = 1, \ldots, n = 20$, at different levels of sparsity ($\pi \in \{0, .5, .9\}$). The second and third rows plot the average maximum λ and minimum λ for which the minimum SURE is achieved.

values selected (minimum and maximum, respectively, for which the SURE criterion is minimized). In these plots, the SCAD and HPLASSO/MCP estimators have nearly identical AERs except in cases of extreme sparsity, where the HPLASSO/MCP estimators have lower AERs. However, neither the empirical Bayes selection nor the SURE selection of λ clearly dominates the other, as none of the smoothed curves lie complete above or below zero in Figure 6.14 (top).

Taken as a whole, the SURE and GCV criteria are not particularly informative in terms of the selection of parameter a, especially when jointly minimized with λ in the case of SURE (strategy (i.)) It is worth emphasizing that while the SURE unbiasedly

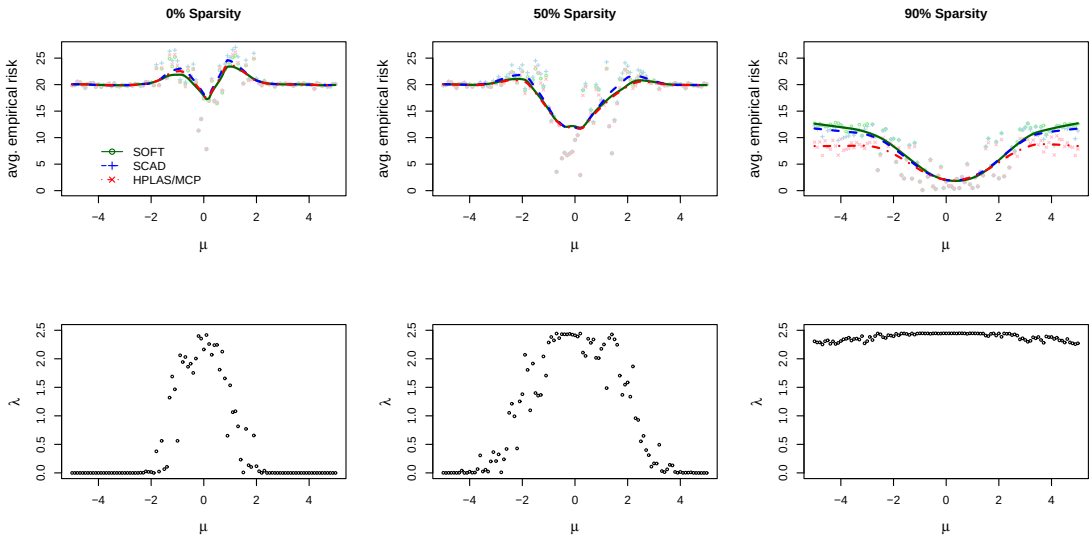

Figure 6.13: Average empirical risks for data-adaptive λ and fixed $a = 3.7$ (strategy (v.)), assuming $z_i \sim N(\theta_i, 1)$, $\theta_i \sim (1-\pi)N(\mu, 1) + \pi\delta_0(\mu)$, $i = 1, \ldots, n = 20$, at different levels of sparsity ($\pi \in \{0, .5, .9\}$). The second row plots the average empirical Bayes-selected λ for each μ.

estimates the true risk, the estimate for a given dataset may be too noisy and too far from the true risk (see, for example, Figure 6.6) for informative tuning parameter selection.

However, with the empirical Bayes estimate of λ, the performance of the SURE- and GCV- selected a was quite similar. This is most likely due to the connection of GCV and SURE documented in Li (1985). While they were indeed similar, the GCV-selected parameters slighted edged out the SURE-selected parameters in terms of AER, especially for small values of μ for the low and moderate sparsity levels. Recall the maximum a was set at ten; subsequent analysis revealed that both criteria could be equivalently minimized at higher values of a, if allowed. Additionally, the (average) maximum a was reported/plotted due to the fact that in many situations, multiple a values (as small as 1.01 and 2.01 for MCP and SCAD, respectively) provided the same minimum criterion.

Thus, it seems that fixing a and then selecting λ separately is the best option. This is illustrated for SURE in Figure 6.14 (bottom), where the AERs are predominantly lower

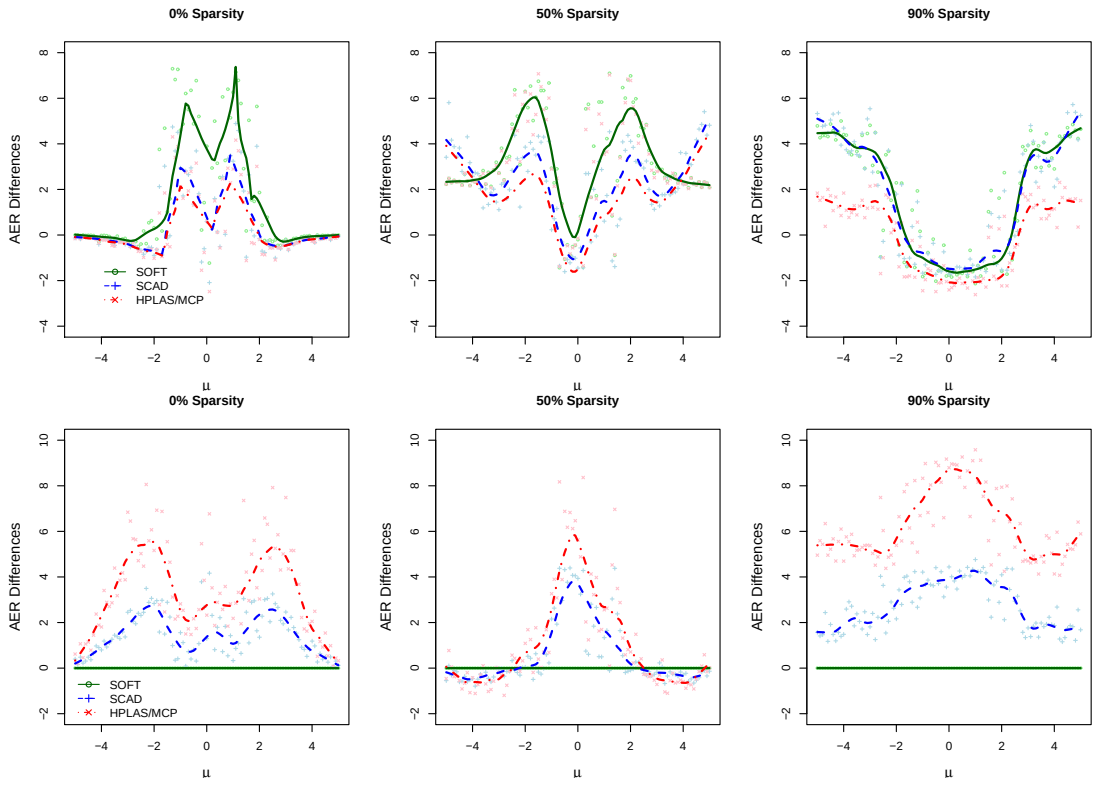

Figure 6.14: Differences in Average Empirical Risks (AER). Top: Difference is 'AER (v.) minus AER (iv.)', so values greater than zero indicate SURE-based λ have lower AER than empirical Bayes-based tuning parameters. Bottom: Difference is 'AER (i.) minus AER (iv.)', so smoothed curves greater than zero indicate SURE-based λ with fixed a have lower AER than SURE-based λ and a parameters.

for strategy (iv.) than for strategy (i). Indeed, Zhang (2010) considers a fixed a parameter in his simulations while Mazumder et al. (2009) select a, but do so using, among other things, a degrees of freedom calibration. Notably, however, neither consider using SURE minimization as a tuning parameter selection procedure. The Johnstone and Silverman (2004) method of selecting λ for fixed a does a remarkable job in terms of small AER, and is comparable to the SURE-based selection of λ, with neither approach clearly dominating the other (Figure 6.14, top).

Finally, we make AER comparisons across thresholding estimators. Strategy (i.), in which both λ and a were selected to minimize the SURE criterion, appears to have the opposite AER performance than the rest of the strategies. The AERs for the

HPLASSO/MCP estimators were generally higher than those for the SCAD and SOFT estimators at the three sparsity levels considered. We attribute the disparity to the relative insensitivity of the SURE criterion when both tuning parameters are unknown. In general, we found that the estimates for the tuning parameters were less variable for strategy (i.) than for any other strategy, particularly for μ in the range [-2,2]. Obvious exceptions occur at high levels of sparsity, when the universal threshold ($\sqrt{2 \log n}$) is nearly always selected (e.g., for strategies (ii.) and (iii.) using the empirical Bayes estimate of λ). For strategies (ii.) and (iii.) using the empirical Bayes λ, the AERs for the SOFT estimators are typically lower than those for the SCAD and HPLASSO/MCP estimators at the low and moderate sparsity levels, with the HPLASSO/MCP estimators having slightly lower AERs than the SCAD estimators. For strategy (iv.) (fixed a, SURE-selected λ) the AERs for the SOFT estimators are still typically lower than those for the SCAD and HPLASSO/MCP estimators at the low and moderate sparsity levels (except perhaps at the extremes), with the HPLASSO/MCP estimators having very similar (although lower) AERs to the SCAD estimators. Strategy (v.) (fixed a, empirical Bayes λ) interestingly resulted in very similar AERs across the thresholding estimators for the low and moderate sparsity levels. But across strategies (ii.)-(v.), in which only one tuning parameter was selected through criterion minimization, the AERs for the HPLASSO/MCP estimators are generally lower than (or equivalent to) those for the SCAD and SOFT estimators at the high sparsity level. This is not unexpected, as the MCP estimator has been shown previously in other contexts (e.g., Zhang, 2010; Mazumder et al., 2009) to perform well under sparsity.

CHAPTER 7

DISCUSSION

This dissertation explored three topics related to penalized estimation and variable selection. First, a versatile and general algorithm, MIST, was proposed for dealing with a wide variety of nonsmoothly penalized objective functions, including but not limited to all presently popular combinations of data fidelity and penalty functions. The MIST algorithm utilizes a judicious choice of majorization to generate a MM algorithm that applies soft-thresholding componentwise and which, in certain settings, allows one to minimize the majorizing function in a single iteration. A suitable convergence theory was established, as well as new results on the convergence of general MM algorithms. Second, the MIST algorithm and theory was extended to the case of finite mixture regression models, with emphasis on linear regression mixtures with unknown common variance. Lastly, a hierarchical motivation was provided for the Minimax Concave Penalty of Zhang (2010), with an investigation of the properties of the resulting univariate thresholding estimator in terms of risk and tuning parameter selection.

An obvious direction for future work includes extending MIST (MIST-MIX) to settings where $p > N$ ($P > N$). Indeed, the simulated examples only consider settings where the total number of observations is greater than the total number of model parameters, in part to ensure the uniqueness condition for the algorithm. While the MIST algorithm has not yet been extensively tested in the $p > N$ ($P > N$) setting, preliminary results show that the algorithm continues to converge and find reasonable solutions when given a reasonable starting value, but tends to converge at a slower rate in comparison with $N > p$ ($N > P$). However, the problem of tuning the algorithm, the development of acceleration procedures and the problem of selecting suitable starting values in situations with multiple local minima, particularly in settings where $p > N$ ($P > N$) but the

number of important predictors $p_0 \ll N$ ($P_0 \ll N$), are all left for further investigation.

The application of MIST-MIX to finite mixtures of GLMs lacking bounded hessians, as well as finite mixtures of censored survival outcomes is currently being explored. For example, the suggestions for Poisson regression presented in Chapter 4 could also be implemented without much difficulty in the finite mixture setting. Censored survival outcomes, however, pose new challenges in the current finite mixture framework and the solutions to these are under development. Also in regard to Chapter 5, it would be nice to find an alternative way to penalize components differentially. This is challenging since we can not postulate group membership of the observations until after the algorithm has converged. While the MIST-MIX "improvement" step rather than "minimization" step for the estimation of $\boldsymbol{\pi}$ does provide some differential penalization across groups, the theory for convergence of such an approach is currently lacking.

In terms of extending the ideas in Chapter 6, it is of interest to explore the use of different prior distributions, and potentially find a hierarchical motivation for the SCAD penalty. The simulations provided in Chapter 6 addressing tuning parameter selection merely scratch the surface. Currently under investigation is the effect of replicate observations using k-fold CV to select tuning parameters. There is an additional consideration with replicate observations and thresholding operators: the order in which the averaging and thresholding operations are performed when estimating $\boldsymbol{\theta}$. Since the thresholding operation is nonlinear, the estimates using the different orderings will generally differ. Tuning parameter selection may also be effected by the choice of ordering.

Finally, as the Johnstone and Silverman (2004) data-adaptive selection of λ provided favorable results in the simulations, it would also be interesting, if possible, to derive a similar marginal likelihood-based approach using the hierarchical structure inherent to HPLASSO to estimate λ (and possibly a). This is an area reserved for future research.

BIBLIOGRAPHY

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000), "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, 403, 503–511.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes.*, Springer-Verlag, New York.

Ando, T., Suguro, M., Kobayashi, T., Seto, M., and Honda, H. (2003), "Multiple fuzzy neural network system for outcome prediction and classification of 220 lymphoma patients on the basis of molecular profiling," *Cancer Sci.*, 94, 906–913.

Annest, A., Bumgarner, R., Raftery, A., and Yeung, K. Y. (2009), "Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data," *BMC Bioinformatics*, 10, 72.

Antoniadis, A., Gijbels, I., and Nikolova, M. (2009), "Penalized likelihood regression for generalized linear models with nonquadratic penalties," *Ann. Inst. Statist. Math.*

Bar, H. Y., Booth, J. G., and Wells, M. T. (2010), "An Empirical Bayes Approach to Variable Selection and QTL Analysis," in *Proceedings of the 25th International Workshop on Statistical Modelling*, ed. A. Bowman.

Becker, M. P., Yang, I., and Lange, K. (1997), "EM algorithms without missing data." *Stat. Methods Med. Res.*, 6, 38–54.

Biernacki, C., Celeux, G., and Govaert, G. (1998), "Assessing a Mixture Model for Clustering with the Integrated Classification Likelihood," Tech. Rep. RR-3521, INRIA.

Binder, H. and Schumacher, M. (2008), "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models," *BMC Bioinformatics*, 9, 14.

Binder, H. and Schumacher, M. (2009), "Incorporating pathway information into boosting estimation of high-dimensional risk prediction models," *BMC Bioinformatics*, 10, 18.

Bohning, D. and Lindsay, B. (1988), "Monotonocity of Quadratic-Approximation Algorithms," *Ann. Inst. Statist. Math.*, 40, 641–663.

Borwein, J. and Lewis, A. (2006), *Convex analysis and nonlinear optimization: theory and examples, Second Edition.*, Canadian Mathematical Society.

Boyd, S. and Vandenberghe, L. (2004), *Convex Optimization.*, Cambridge University Press.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

Breiman, L. and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *J. Amer. Statist. Assoc.*, 80, 580–598.

Cai, T., Huang, J., and Tian, L. (2009), "Regularized estimation for the accelerated failure time model," *Biometrics*, 65, 394–404.

Carvalho, C., Polson, N., and Scott, J. (2010), "The horseshoe estimator for sparse signals," *Biometrika*, to appear.

Chrétien, S. and Hero, A. O. (2008), "On EM algorithms and their proximal generalizations." *ESAIM Journ. on Probability and Statistics*, 12, 308–326.

Clarke, F. H. (1990), *Optimization and Nonsmooth Analysis.*, SIAM, Philadelphia.

Combettes, P. and Wajs, V. (2005), "Signal Recovery by Proximal Forward-Backward Splitting," *Multiscale Model. Simul.*

Cox, D. R. (1972), "Regression models and life-tables (with Discussion)," *J. Roy. Statist. Soc. Ser. B*, 34, 187–220.

Cox, D. R. (1975), "Partial likelihood," *Biometrika*, 62, 269–276.

Datta, S., Le-Rademacher, J., and Datta, S. (2007), "Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO," *Biometrics*, 63, 259–271.

Daubechies, I., Defreise, M., and De Mol, C. (2004), "An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint," *Commun. Pure Appl. Math.*, pp. 1413–1457.

de Leeuw, J. (1994), "Block-relaxation Algorithms in Statistics," *Information systems and data analysis*.

De Mol, C., De Vito, E., and Rosasco, L. (2008), "Elastic-Net Regularization in Learning Theory," *arXiv*.

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. Ser. B*, 39, 1–38.

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

Droge, B. (1993), "On finite sample properties of adaptive least squares regression estimates," *Statistics*, 24, 181–203.

Droge, B. (1998), "Minimax regret analysis of orthogonal series regression estimation: Selection versus shrinkage," *Biometrika*, 85, 631–643.

Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. (2004), "Least angle regression," *Ann. Statist.*, 32, 407–452.

Engler, D. and Li, Y. (2009), "Survival analysis with high-dimensional covariates: an application in microarray studies," *Statist. Appl. Gen. Mol. Biol.*, 8, Article 14.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties." *J. Amer. Statist. Assoc.*

Fan, J. and Li, R. (2002), "Variable Selection for Cox's Proportional Hazards Model and Frailty Model," *Ann. Statist.*, 30, 74–99.

Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2009a), *SIS: Sure Independence Screening*, R package version 0.2.

Fan, J., Samworth, R., , and Wu, Y. (2009b), "Ultrahigh dimensional variable selection: beyond the linear model," *Journal of Machine Learning Research, to appear*.

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Regularization Paths for Generalized Linear Models via Coordinate Descent," Tech. rep., Dept. of Statistics, Stanford University.

Fygenson, M. and Ritov, Y. (1994), "Monotone estimating equations for censored data." *Ann. Statist.*, 22, 732–746.

Gao, H. and Bruce, A. G. (1997), "Waveshrink with firm shrinkage," *Statist. Sinica*, 7, 855–874.

Geman, D. and Reynolds, G. (1992), "Constrained restoration and the recovery of discontinuities." *IEEE Trans. Patt. Anal. Mach. Intell*, 14, 367–383.

Griffin, J. and Brown, P. (2005), "Alternative prior distributions for variable selection with very many more variables than observations." Tech. rep., Dept. of Statistics, University of Warwick.

Griffin, J. and Brown, P. (2007), "Bayesian adaptive lassos with non-convex penalization." Tech. rep., Dept. of Statistics, University of Warwick.

Gruen, B. and Leisch, F. (2008), "FlexMix Version 2: Finite mixtures with concomitant variables and varying and constant parameters," *J. of Statist. Soft.*, 28, 1–35.

Gui, J. and Li, H. (2005a), "Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data," *Bioinformatics*, 21, 3001–3008.

Gui, J. and Li, H. (2005b), "Threshold Gradient Descent Method for Censored Data Regression with Applications in Pharmacogenomics," *Pacific Symposium on Biocomputing*, 10, 272–283.

Hale, E., Yin, W., and Zhang, Y. (2008), "Fixed-point Continuation for l1-minimization: Methodology and Convergence." *SIAM J. Optim.*, 19, 1107–1130.

Hastie, T. and Efron, B. (2007), *lars: Least Angle Regression, Lasso and Forward Stagewise*, R package version 0.9-7.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (1993), *Convex Analysis and Minimization Algorithms I: Fundamentals.*, Springer.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (1996), *Convex Analysis and Minimization Algorithms I: Fundamentals.*, Springer.

Huang, J., Ma, S., and Xie, H. (2006), "Regularized estimation in the accelerated failure time model with high-dimensional covariates," *Biometrics*, 62, 813–820.

Hunter, D. and Li, R. (2005), "Variable Selection Using MM Algorithms," *Ann. Statist.*, 33, 1617–1643.

Johnson, B. A., Lin, D. Y., and Zeng, D. (2008), "Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models," *J. Amer. Statist. Assoc.*, pp. 672–680.

Johnson, L. and Strawderman, R. (2009), "Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data." *Biometrika*.

Johnstone, I. M. and Silverman, B. W. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *Ann. Statist.*, 32, 1594–1649.

Johnstone, I. M. and Silverman, B. W. (2005), "Empirical Bayes Selection of Wavelet Thresholds," *Ann. Statist.*, 33, 1700–1752.

Khalili, A. and Chen, J. (2007), "Variable Selection in Finite Mixture of Regression Models," *J. Amer. Statist. Assoc.*, 102, 1025–1038.

Kim, Y., Choi, H., and Oh, H.-S. (2008), "Smoothly Clipped Absolute Deviation on High Dimensions," *J. Amer. Statist. Assoc.*, 103, 1665–1673.

Lange, K. (1995), "Optimization Transfer Using Surrogate Objective Functions," *J. Roy. Statist. Soc. Ser. B*, 57, 425–437.

Lange, K. (2004), *Optimization.*, Springer, New York, USA.

Lange, K., Hunter, D., and Yang, I. (2000), "Optimization Transfer Using Surrogate Objective Functions," *J. of Comp. Graph. Statist.*, 9, 1–20.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006), *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood.*, Monographs on Statistics and Applied Probability 106, Chapman and Hall, Boca Raton, Florida.

Li, H. and Gui, J. (2004), "Partial Cox regression analysis for high-dimensional microarray gene expression data," *Bioinformatics*, 20 Suppl 1, i208–215.

Li, H. and Luan, Y. (2005), "Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data," *Bioinformatics*, 21, 2403–2409.

Li, K.-C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation," *Annals of Statist.*, 13, 1352–1377.

Li, K.-C. and Hwang, J. T. (1984), "The Data-Smoothing Aspect of Stein Estimates," *Annals of Statist.*, 12, 887–897.

Luenberger, D. G. and Ye, Y. (2008), *Linear and nonlinear programming.*, International Series in Operations Research & Management Science, 116, Springer, New York, third edn.

Ma, S. and Huang, J. (2007), "Additive risk survival model with microarray data," *BMC Bioinformatics*, 8, 192.

Mäkelä, M. and Neittaanmaäki, P. (1992), *Nonsmooth optimization: analysis and algorithms with applications to optimal control.*, World Scientific, New Jersey.

Martinez-Climent, J. A., Alizadeh, A. A., Segraves, R., Blesa, D., Rubio-Moscardo, F., Albertson, D. G., Garcia-Conde, J., Dyer, M. J., Levy, R., Pinkel, D., and Lossos, I. S. (2003), "Transformation of follicular lymphoma to diffuse large cell lymphoma is associated with a heterogeneous set of DNA copy number and gene expression alterations," *Blood*, 101, 3109–3117.

Mazumder, R., Friedman, J., and Hastie, T. (2009), "Sparsenet: Coordinate descent with non-convex penalties," Tech. rep., Department of Statistics, Stanford University.

McLachlan, G. J. and Krishnan, T. (2008), *The EM Algorithm and Extensions.*, Wiley-Interscience, 2 edn.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models.*, Wiley-Interscience.

Melkonyan, T. (2010), "Generalized Derivatives," Tech. rep., University of Nevada, Reno, Available: http://www.cabnr.unr.edu/melkonyan/lecture%20notes.aspx.

Meyer, R. R. (1976), "Sufficient conditions for the convergence of monotonic mathematical programming algorithms," *J. Comput. System. Sci.*, 12, 108–121.

Meyer, R. R. (1977), "A comparison of the forcing function and point-to-set mapping approaches to convergence analysis," *SIAM J. Control Optimization*, 15, 699–715.

Nikolova, M. (2000), "Local strong homogeneity of a regularized estimator," *SIAM J. Appl. Math.*, 61, 633–658.

Ortega and Rheinboldt (2000), *Iterative Solution of Nonlinear Equations in Several Variables.*, Society for Industrial and Applied Mathematics, New York.

Park, M. Y. and Hastie, T. (2007), "L1-regularization path algorithm for generalized linear models," *J. Roy. Statist. Soc. Ser. B*, 69, 659–677.

Park, T. and Casella, G. (2008), "The Bayesian Lasso," *J. Amer. Statist. Assoc.*, 103, 681–686.

Polak, E. (1987), "On the mathematical foundations of nondifferentiable optimization in engineering design," *SIAM Rev.*, 29, 21–89.

R Development Core Team (2005), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Roland, C. and Varadhan, R. (2005), "New iterative schemes for nonlinear fixed point problems, with applications to problems with bifurcations and incomplete-data problems," *Appl. Numer. Math.*, 55, 215–226.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., Lpez-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., and Staudt, L. M. (2002), "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma," *N. Engl. J. Med.*, 346, 1937–1947.

Rosset, S. and Zhu, J. (2007), "Piecewise linear regularized solution paths," *Ann. Statist.*, 35, 1012–1030.

Saeki, K., Miura, Y., Aki, D., Kurosaki, T., and Yoshimura, A. (2003), "The B cell-specific major raft protein, Raftlin, is necessary for the integrity of lipid raft and BCR signal transduction," *EMBO J.*, 22, 3015–3026.

Schifano, E. D. (2007), "Generalized Wavelet Thresholding: Estimation and Hypothesis Testing with Applications to Array Comparative Genomic Hybridization." Master's thesis, Cornell University.

Sha, N., Tadesse, M. G., and Vannucci, M. (2006), "Bayesian variable selection for the analysis of microarray data with censored outcomes," *Bioinformatics*, 22, 2262–2268.

She, Y. (2009), "Thresholding-based iterative selection procedures for model selection and shrinkage," *Electron. J. Stat.*, 3, 384–415.

Sinisi, S. E., Neugebauer, R., and van der Laan, M. J. (2006), "Cross-validated bagged prediction of survival," *Statist. Appl. Gen. Mol. Biol.*, 5, Article12.

Sohn, I., Kim, J., Jung, S. H., and Park, C. (2009), "Gradient lasso for Cox proportional hazards model," *Bioinformatics*, 25, 1775–1781.

Städler, N., Buhlmann, P., and van de Geer, S. (2010), "l1-Penalization for Mixture Regression Models," *TEST, to appear*.

Stein, C. (1981), "Estimation of the mean of a multivariate normal mean," *Annals of Statist.*, 9, 1135–1151.

Strawderman, R. L. and Wells, M. T. (2010), "On Hierarchical Prior Specifications and Penalized Likelihood," *IMS Collections*, to appear.

Strawderman, W. E. (1971), "Proper Bayes minimax estimators of the normal multivariate normal distribution," *Ann. Math. Statist.*, 42, 385–388.

Takada, Y. (1979), "Stein's positive part estimator and bayes estimator," *Ann. Inst. Statist. Math.*, 31, 177–183.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso." *J. Roy. Statist. Soc. Ser. B*, pp. 267–288.

Tibshirani, R. (1997), "The lasso method for variable selection in the Cox model," *Statist. in Med.*, 16, 385–395.

Tibshirani, R. J. (2009), "Univariate shrinkage in the cox model for high dimensional data," *Statist. Appl. Gen. Mol. Biol.*, 8, Article21.

Tseng, P. (2001), "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *J. Optim. Theory Appl.*, 109, 475–494.

Tseng, P. (2004), "An analysis of the EM algorithm and entropy-like proximal point methods," *Mathematics of Operations Research*, 29, 27–44.

Vaida, F. (2005), "Parameter convergence for EM and MM algorithms," *Statist. Sinica*, 15, 831–840.

Varadhan, R. and Roland, C. (2008), "Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm," *Scand. J. Statist.*, 35, 335–353.

Wang, S., Nan, B., Zhu, J., and Beer, D. G. (2008), "Doubly penalized buckley-james method for survival data with high-dimensional covariates," *Biometrics*, 64, 132–140.

Wasserman, L. (2006), *All of Nonparametric Statistics.*, Springer Texts in Statistics, Springer, New York.

Wu, C.-F. J. (1983), "On the convergence properties of the EM algorithm," *Ann. Statist.*, 11, 95–103.

Zangwill, W. I. (1969), *Nonlinear Programming; a Unified Approach.*, Prentice-Hall International Series in Management, Englewood Cliffs, N.J.

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *Ann. Statist.*, 38, 894–942.

Zhang, D. and Zhang, M. (2007), "Bayesian profiling of molecular signatures to predict event times," *Theor. Biol. Med. Model.*, 4, 3.

Zhang, H. H. and Lu, W. (2007), "Adaptive LASSO for Cox's proportional hazards model," *Biometrika*, 94, 691–703.

Zhang, M., Zhang, D., and Wells, M. T. (2010a), "Generalized thresholding estimators for high-dimensional location parameters," *Statist. Sinica*, 20, 911–926.

Zhang, Y., Li, R., and Tsai, C.-L. (2010b), "Regularization Parameter Selections via Generalized Information Criterion," *J. Amer. Statist. Assoc.*, 105.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *J. Amer. Statist. Assoc.*, 101, 1418–1429.

Zou, H. and Hastie, T. (2005), "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc. Ser. B*, pp. 301–320.

Zou, H. and Hastie, T. (2008), *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*, R package version 1.0-5.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Statist.*, 36, 1509–1533.

Zou, H. and Zhang, H. H. (2009), "On the adaptive elastic-net with a diverging number of parameters," *Ann. Statist.*