RICE UNIVERSITY

Parametric Classification and Variable Selection by the Minimum Integrated Squared Error Criterion

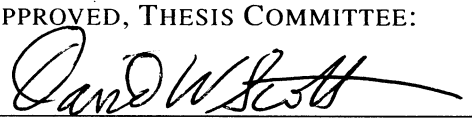
by

Eric C. Chi

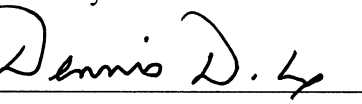
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

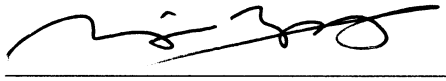
APPROVED, THESIS COMMITTEE:



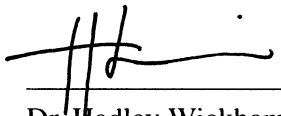
Dr. David W. Scott, *Chair*
Noah Harding Professor of Statistics, Rice University



Dr. Dennis D. Cox
Professor of Statistics, Rice University



Dr. Yin Zhang
Professor of Computational and Applied Mathematics, Rice University



Dr. Hadley Wickham
Assistant Professor of Statistics, Dobelman Family Junior Chair, Rice University

HOUSTON, TEXAS
AUGUST 2011

ABSTRACT

Parametric Classification and Variable Selection by the Minimum Integrated Squared Error Criterion

by

Eric C. Chi

This thesis presents a robust solution to the classification and variable selection problem when the dimension of the data, or number of predictor variables, may greatly exceed the number of observations. When faced with the problem of classifying objects given many measured attributes of the objects, the goal is to build a model that makes the most accurate predictions using only the most meaningful subset of the available measurements. The introduction of ℓ_1 regularized model fitting has inspired many approaches that simultaneously do model fitting and variable selection. If parametric models are employed, the standard approach is some form of regularized maximum likelihood estimation. While this is an asymptotically efficient procedure under very general conditions, it is not robust. Outliers can negatively impact both estimation and variable selection. Moreover, outliers can be very difficult to identify as the number of predictor variables becomes large.

Minimizing the integrated squared error, or L_2 error, while less efficient, has been shown to generate parametric estimators that are robust to a fair amount of contamination in several contexts. In this thesis, we present a novel robust parametric regression model for the binary classification problem based on L_2 distance, the logistic L_2 estimator (L_2E). To perform simultaneous model

fitting and variable selection among correlated predictors in the high dimensional setting, an elastic net penalty is introduced. A fast computational algorithm for minimizing the elastic net penalized logistic L_2E loss is derived and results on the algorithm's global convergence properties are given. Through simulations we demonstrate the utility of the penalized logistic L_2E at robustly recovering sparse models from high dimensional data in the presence of outliers and inliers. Results on real genomic data are also presented.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my thesis advisor, David Scott. I could not have asked for a better mentor. His curiosity, creativity, energy, and patience provided a very necessary steadying influence as I made the transition from being a consumer of results to producing results of my own. It has been a real privilege to learn from David how to do research, and I hope there will be many occasions in the future for me to pick up more pearls from him. But the most important lessons David's taught me have little do with the research and more to do with how to conduct oneself with people. It is no surprise that David is so well liked and trusted by anybody who has had any serious interaction with him. He just treats people right. I can only hope my career will be similarly marked by both quality work and personal integrity.

I also owe a great deal of thanks to many others for helping me grow as a researcher during my Ph.D. training. I thank my committee members for their feedback on my thesis and all the recommendations and advice they gave me to improve the quality of my work. I also thank Tammy Kolda for her mentorship while I was an intern at Sandia National Labs. Even though her expertise is outside of statistics I am without a doubt a better statistician because of her questions, advice, critique, and feedback from my summer there. I also thank my peers in the program; our impromptu conversations in the hall-way often sharpened my understanding of subtleties in statistics that I would not have resolved on my own. And of course I also want to thank them for their friendship. It meant a lot to me whenever I was asked how Remy and Jocelyn were

doing. I especially want to thank Terrence Savitsky, Alejandro Marcelo-Cruz, and Garrett Grolemund for being the kind of traveling companions one needs on the path to a Ph.D. Additionally, I would like to thank Dr. Michael Keyton, my highschool geometry teacher, for sharing his delight in mathematical elegance with me. If I had to blame a single person for how I ended up on this wonderful path, it would be him.

My family deserves a very special thanks. I would like to thank my parents for their support during my long tenure as a student and my wonderful in-laws for all their enthusiastic cheerleading and moral support. I thank my precious son, Jeremiah, for teaching me to be ten times more efficient with my waking hours but also for reminding me not to take myself or my work too seriously. I thank my best friend and lovely wife, Jocelyn, for her unending support and unwavering belief in me. There are many hard decisions that I could not have made without her.

Finally, I thank the living creator God for the blessing of getting to do the work I do. We can think and comprehend so that we might better know him. My hope is that this training will be used fully to that end. Soli Deo Gloria.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Regression, classification and variable selection problems in high dimensional data are becoming routine in fields ranging from finance to genomics. In the latter case, technologies such as expression arrays and SNP chips have made it possible to comprehensively query a patient's genetic profile and transcriptional activity [59, 25]. Patterns in these profiles can help refine subtypes of a disease according to sensitivity to treatment options or identify previously unknown genetic components of a disease's pathogenesis.

The immediate statistical challenge is finding those patterns when the number of predictors far exceeds the number of samples. To that end the Least Absolute Shrinkage and Selection Operator (LASSO) has been quite successful at addressing "the small n, big p problem" [11, 51, 43, 52, 10]. Indeed, ℓ_1 penalized maximum likelihood model fitting has inspired many related approaches that simultaneously do model fitting and variable selection. While these developments have been important first steps, noticeably less attention has been given to extending these methods to handle model misspecification, e.g. contamination or heterogeneity in the data.

Being able to deal with contaminants in a principled manner becomes important as outliers and inliers can become harder to detect as the dimensionality of the predictor space increases. This is a material issue, because, as will be shown in this thesis, outliers and in-

liers can seriously hamper the variable selection procedure when LASSO like approaches are taken. This thesis presents a response to dealing with fitting parametric models when $n \ll p$ and statistical robustness is desired. Specifically we present a regularized minimum distance estimator for fitting parametric models data with binary responses and high dimensional covariates.

1.1 Notation, Preliminaries, and Problem Formulation

Throughout this thesis we will use the following notation. Random variables are denoted by upper case letters, e.g. X. Vectors are denoted in lowercase boldface letters, e.g. $\mathbf{b}$. All vectors are to be taken as column vectors. The ith element of a vector $\mathbf{x}$ is denoted x_i. The ℓ_q norm of a vector $\mathbf{a} \in \mathbb{R}^n$ for $q > 0$ is denoted by $\|\mathbf{a}\|_q$ and is defined to be

$$\|\mathbf{a}\|_q = \left(\sum_{i=1}^{n} |a_i|^q \right)^{\frac{1}{q}}.$$

All matrices are denoted in uppercase boldface letters, e.g. $\mathbf{A}$. If $\mathbf{A}$ is a matrix, we denote the transpose of its ith row by $\mathbf{a}_i$ and its ijth element by a_{ij}.

In the supervised learning problem, the goal is to identify a model that accurately and concisely summarizes the relationship between a collection of covariates in the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\mathsf{T} \in \mathbb{R}^{n \times p}$ and responses $\mathbf{y} \in \mathbb{R}^n$. Our concern is finding optimal parametric models, specifically parametric models of the density of $\mathbf{y}$ that are functions of linear predictors $\mathbf{X}\boldsymbol{\theta}$ of the covariates where $\boldsymbol{\theta} \in \mathbb{R}^p$. Optimality is assessed with respect to a combination of two quantities: lack of fit and model complexity. Lack of fit is measured by a nonnegative loss function $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$. Model complexity is measured by a nonnegative penalty function $J : \mathbb{R}^p \to \mathbb{R}_+$, with $J(\mathbf{0}) = 0$. We wish to find

$$\hat{\boldsymbol{\theta}}(\lambda) = \arg\min_{\boldsymbol{\theta}} L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta}). \tag{1.1}$$

where $\lambda \geq 0$ is a regularization parameter that controls the tradeoff between model fit and model complexity.

1.1.1 Maximum Likelihood Estimation

A common choice for L when parametric models are considered is the negative log-likelihood. If this choice is made for L with $\lambda = 0$ then the minimizing parameter is then a maximum likelihood estimator (MLE). Often this choice is justified on the grounds that maximum likelihood estimation is an asymptotically efficient procedure under very general conditions [50]. Additionally the negative log-likelihood of many important distributions are convex, expediting computation of the MLE. For example the negative log-likelihood of any member of the exponential family is convex. Consider the likelihood of a member of the exponential family with natural parameter η

$$P(\mathbf{z}|\boldsymbol{\eta}) = e^{\mathbf{z}^\mathsf{T}\boldsymbol{\eta} - G(\boldsymbol{\eta})} P_0(\mathbf{z}),$$

where the cumulant function $G(\boldsymbol{\eta}) = \log \int \exp(\mathbf{z}^\mathsf{T}\boldsymbol{\eta}) P_0(\mathbf{z}) d\mathbf{z}$ ensures that $P(\mathbf{z}|\boldsymbol{\eta})$ integrates to unity. Note that G is a convex function of η. Therefore, $-\log P(\mathbf{z}|\boldsymbol{\eta})$ is convex since it is the sum of a convex function and an affine function of η. If $L(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2^2$, which is proportional to the negative log-likelihood of one of the most celebrated members of the exponential family, the normal distribution with known variance, and $\lambda = 0$, then (1.1) becomes the ubiquitous ordinary least squares regression problem.

1.1.2 When $n < p$

It is now common to be confronted with data where the number of possible predictive features p far exceeds the number of training samples, n. Consider a typical pharmacogenomics question: given cohort of breast cancer patients, a fraction of which responds well

to a chemotherapeutic regimen, is there a gene expression pattern that can be used to explain sensitivity to said regimen in patients? Unfortunately when $n < p$, maximum likelihood estimation becomes unstable in the sense that a small perturbation in the data can result in a disproportionate variation in the fitted model. An unstable estimation procedure has high variance. Consider the least squares problem

$$\hat{\boldsymbol{\theta}}(\lambda = 0) = \arg\min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2,$$

when $n < p$. The rank of $\mathbf{X}$ is at most n and therefore $\mathbf{X}$ has a non-empty null space. The upshot is that there are infinitely many $\boldsymbol{\theta}$ such that $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$, and for any such minimizing choice of $\boldsymbol{\theta}$ the model is said to "overfit" the data because the resulting $\boldsymbol{\theta}$ is essentially capturing the random variation of the particular data set used to fit the model instead of the systematic variation that exists in the population. The resulting model will very likely have poor predictive performance on new observations drawn from the population.

A standard strategy for stabilizing an estimation procedure is regularization, i.e. taking $\lambda > 0$ in (1.1). To understand why regularization helps stabilize estimation recall that the mean squared error of an estimator $\hat{\boldsymbol{\theta}}$ of a parameter $\boldsymbol{\theta}$ can be expressed as the following sum

$$\text{MSE}(\hat{\boldsymbol{\theta}}) = E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}]^2 = E[\hat{\boldsymbol{\theta}} - E\hat{\boldsymbol{\theta}}]^2 + (E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}])^2,$$

where the variance of $\hat{\boldsymbol{\theta}}$ is the first term on the right hand side and the bias of $\hat{\boldsymbol{\theta}}$ is the second term. If $J \to \infty$ in (1.1) whenever $|\theta_i| \to \infty$ for all i, we see that as λ increases the estimator $\hat{\boldsymbol{\theta}}$ is progressively shrunk towards $\mathbf{0}$. With respect to the MSE decomposition, variance is decreased at the expense of bias. For sufficiently large λ, there is no variance as $\hat{\boldsymbol{\theta}} = 0$ and the MSE is completely due to estimator bias. For some intermediate $\lambda > 0$ the MSE is minimal. Since $\hat{\boldsymbol{\theta}}$ shrinks to zero as the penalty increases, the resulting estimators are commonly known as shrinkage estimators.

1.1.3 Choosing the penalty function J

We restrict our attention to penalty functions that are separable in the absolute value of their arguments, i.e. $J(\boldsymbol{\theta}) = \sum_{j=1}^{p} \phi_j(|\theta_j|)$ for univariate functions $\phi_j : \mathbb{R}_+ \to \mathbb{R}$. A wide variety of separable choices for J can be summarized by the class of univariate penalty functions that have the form

$$\phi_j(z) = \alpha_1 z^q + \alpha_2 z^r \tag{1.2}$$

where $\alpha_i \geq 0$. Best subset selection corresponds to $\alpha_2 = 0$ and $q = 0$. The LASSO corresponds to $\alpha_2 = 0$ and $q = 1$ [52]. Ridge regression corresponds to $\alpha_1 = 0$ and $q = 2$ [26]. The LASSO and ridge regression are special cases of the Elastic Net which is a weighted mix of both penalties by setting $q = 1$ and $r = 2$ [62]. Bridge regression is defined by the set of penalties for which $\alpha_2 = 0$ and $q \geq 0$ and so the LASSO and best subset selection are special cases of it [20].

Bridge penalties for which $q < 1$ are concave and penalties for which $q \leq 1$ are not differentiable at the origin. Penalties that are concave, like the Bridge penalties with $q < 1$ and the SCAD penalty [19], produce less biased estimates [23] but create computationally more challenging optimization problems compared to their convex counterparts when the accompanying loss function is convex. Such combinations define convex programs for which exist a body of well developed solvers based on interior point methods [7].

Univariate penalty functions ϕ_j that are differentiable everywhere except at the origin like the LASSO, Elastic Net, SCAD, and concave Bridge penalties perform variable selection. These penalties incentivize sparse models, models for which many elements of $\hat{\boldsymbol{\theta}}$ are zero. These regularization penalties are preferred when the true set of important features is suspected to be only a small fraction of the set of measured features.

These notions of prior belief can be formalized in a Bayesian setting. With a least squares loss, ℓ_1 regularized minimization corresponds to picking the posterior mode of a normal likelihood subject to a prior belief in a sparse model formalized by a Laplacian

prior on θ. Conversely, if it is suspected that many features are making small contributions to the response variable, ridge regression will yield the better model. This prior belief would be formalized by a Bayesian as a Normal prior on θ. Moreover, ridge regression is more stable with respect to variable selection in the presence of correlation, including and excluding groups of correlated variables in the fitted model. Given two correlated variables the LASSO will tend to select either one or the other, while ridge regression will select both. See [62] for a nice demonstration for how the ridge penalty accomplishes this by "decorrelating" correlated variables. The Elastic Net penalty with $\alpha_1 > 0$ occupies the unique position of inducing variable selection and being convex making it an ideal candidate for pairing with a convex loss function, such as a negative log-likelihood from the exponential family. For the rest of this thesis, since our work is motivated by genomic data which is known to have correlated covariates, we will focus on the Elastic Net penalty because it produces sparse models but includes and excludes groups of correlated variables. In the next section we motivate the need for a LASSO-type penalty like the Elastic Net by characterizing the performance of a naive classification procedure in the $n \ll p$ setting.

1.2 A Case for Variable Selection: Nearest Centroids

Let $\{(X_i, Y_i)\}$ be i.i.d. draws from a $2p$-dimensional multivariate normals with identity covariance matrix and $X_i, Y_i \in \mathbb{R}^p$ are samples from two distinct clusters. Let the mean be $(\mu_1, 0, \ldots, 0, \mu_2, 0, \ldots, 0)$. That is the mean of X_i is $(\mu_1, 0, \ldots, 0)$ and similarly the mean of Y_i is $(\mu_2, 0, \ldots, 0)$. Let $\mathbf{e}_1 \in \mathbb{R}^p$ with its first component 1 and the rest 0. Let $\Delta = \mu_1 - \mu_2$. The best way to distinguish points of one cluster from points of the other cluster is by looking at their values in the first coordinate. Not knowing the true distribution of $\{(X_i, Y_i)\}$ in advance, however, one might be tempted to use their projections onto the difference of centroids vector $\bar{\mathbf{x}} - \bar{\mathbf{y}} = (\bar{x}_1 - \bar{y}_1, \ldots, \bar{x}_p - \bar{y}_p)$ where $\bar{x}_j$ denotes the jth component of the centroid of the X_is. So, a new observed point would be classified to the

cluster that had the nearest centroid in Euclidean distance. The following proposition shows the futility of using the difference in centroids for discrimination as p increases for fixed n. Specifically, let α_p denote the angle between $\mathbf{e}_1$ and the difference in sample means. Then as p increases, $\alpha_p \longrightarrow \pi/2$. The difference in sample means becomes orthogonal to the direction which contains all the discriminatory power as p increases.

Proposition 1.2.0.1. *Let* $\cos\alpha_p = \mathbf{e}_1^{\mathsf{T}}(\bar{\mathbf{x}} - \bar{\mathbf{y}})/\|\bar{\mathbf{x}} - \bar{\mathbf{y}}\|_2$ *then*

$$\alpha_p \sim AN\left(\arccos\left(\sqrt{\frac{n}{2p}}\Delta\right), \frac{1 + \frac{n}{4p}\Delta^2}{p - \frac{n}{2}\Delta^2}\right)$$

Proof.

$$\cos\alpha_p = \frac{Z_1}{\sqrt{\sum_{k=1}^{p} Z_k^2}}$$

where $Z_i \sim N(0, 2/n)$ for $i > 1$ and $Z_1 \sim N(\Delta, 2/n)$. Note that $W_{p-1} = \frac{n}{2}\sum_{k=2}^{p} Z_k^2 \sim \chi^2_{p-1}$. Therefore,

$$\cos\alpha_p \stackrel{d}{=} \frac{V_1}{\sqrt{V_1^2 + W_{p-1}}}$$

where $V_1 \sim N(\sqrt{\frac{n}{2}}\Delta, 1)$. Note that by the SLLN $W_{p-1} \longrightarrow \infty$ a.s. as p increases. We suspect then that as p grows $V_1^2 + W_{p-1}$ approaches W_{p-1}. We now show that

$$\frac{\sqrt{V_1^2 + W_{p-1}}}{\sqrt{W_{p-1}}} \xrightarrow{p} 1.$$

Fix $\epsilon > 0$.

$$
\begin{aligned}
P\left(\left|\frac{V_1^2+W_{p-1}}{W_{p-1}}-1\right|>\epsilon\right) &= P(V_1^2>\epsilon W_{p-1}) \\
&= E[P(V_1^2>\epsilon W_{p-1}|W_{p-1})] \\
&\le E\left[\frac{E(V_1^2|W_{p-1})}{\epsilon W_{p-1}}\right] \qquad \text{by Markov inequality} \\
&= E\left[\frac{E(V_1^2)}{\epsilon W_{p-1}}\right] \qquad \text{by independence of } V_1 \text{ and } W_{p-1} \\
&= E\left[\frac{\frac{n}{2}\Delta^2+1}{\epsilon W_{p-1}}\right] \\
&= \frac{\frac{n}{2}\Delta^2+1}{\epsilon}\int_0^\infty \frac{w^{\frac{p-1}{2}-1-1}}{2^{\frac{p-1}{2}}\Gamma(\frac{p-1}{2})}e^{-\frac{w}{2}}dw \\
&= \frac{\frac{n}{2}\Delta^2+1}{\epsilon}\frac{2^{\frac{p-1}{2}-1}}{2^{\frac{p-1}{2}}}\frac{\Gamma(\frac{p-1}{2}-1)}{\Gamma(\frac{p-1}{2})} \\
&= \frac{\frac{n}{2}\Delta^2+1}{\epsilon}\frac{1}{2}\frac{\Gamma(\frac{p-1}{2}-1)}{\frac{p-1}{2}\Gamma(\frac{p-1}{2}-1)} \\
&= \frac{\frac{n}{2}\Delta^2+1}{\epsilon(p-1)}.
\end{aligned}
$$

So,

$$
\frac{V_1}{\sqrt{V_1^2+W_{p-1}}} = \frac{V_1}{\sqrt{W_{p-1}}}\times\frac{\sqrt{W_{p-1}}}{\sqrt{V_1^2+W_{p-1}}} \xrightarrow{p} \frac{V_1}{\sqrt{W_{p-1}}}.
$$

But $\xi_p = V_1/\sqrt{W_{p-1}} = (\sqrt{p-1})^{-1}T_p$ where T_p is distributed as a non-central t with non-centrality parameter $\delta = \sqrt{n/2}\Delta$ and degrees of freedom $\nu = p$. T_p, however, is asymptotically normal [28] with mean δ and standard deviation $\left(1+\frac{\delta^2}{2\nu}\right)^{\frac{1}{2}}$. Therefore,

$$
\sqrt{p}\left(\xi_p - \sqrt{\frac{n}{2p}}\Delta\right) \xrightarrow{d} N\left(0, 1+\frac{\Delta^2 n}{4p}\right)
$$

Recall that $\frac{d}{dx}\arccos(x) = -(1-x^2)^{-1/2}$. By the Delta method

$$\sqrt{p}\left(\arccos(\xi_p) - \arccos(\sqrt{\frac{n}{2p}}\Delta)\right) \xrightarrow{d} N\left(0, \frac{1+\frac{n}{4p}\Delta^2}{1-\frac{n}{2p}\Delta^2}\right).$$

So, for p very large,

$$\alpha_p \sim N\left(\arccos(\sqrt{\frac{n}{2p}}\Delta), \frac{1+\frac{n}{4p}\Delta^2}{p-\frac{n}{2}\Delta^2}\right)$$

□

1.3 Binary Response Data

Suppose in our supervised learning problem the responses are binary $y_i \in \{0,1\}$. Binary responses arise in many fields. For example, in epidemiology the response indicates disease status and the associated feature vector contains potential risk factors for the disease under study. Given new and unlabeled observations, one goal is to accurately predict their appropriate labels. In other words, we want to accurately estimate the conditional probability, $P(Y_i = 1|X_i = \mathbf{x}_i)$. The standard approach to estimating the conditional probability is to fit a logistic model by maximum likelihood estimation [38]. In logistic regression the mean response is modeled as a function of a linear combination of the features

$$P(Y_i = 1|X_i = \mathbf{x}_i) = F(\theta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta}), \tag{1.3}$$

where $F(u) = 1/(1+\exp(-u))$. The parameter $(\theta_0, \boldsymbol{\theta}) \in \mathbb{R}^{p+1}$ is then determined by maximizing the likelihood function

$$(\hat{\theta}_0, \hat{\boldsymbol{\theta}}) = \arg\max_{\theta_0,\boldsymbol{\theta}} \prod_{i=1}^{n} F(\theta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta})^{y_i} \left(1 - F(\theta_0 + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\theta})\right)^{1-y_i}. \tag{1.4}$$

This logistic model is equivalent to setting $Y_i = I(Y_i^* > 0)$ where $Y_i^* = \theta_0 + X_i^{\mathsf{T}}\boldsymbol{\theta} + \epsilon_i$ are latent response variables with the ϵ_i as i.i.d. perturbations with distribution $F(u) = 1/(1 + \exp(-u))$.

Another goal is to identify those features that have the greatest partial correlation with the response variables for inclusion into the model and conversely exclude features that have very little to do with the response variables. This is especially of interest when $p \gg n$, but is still of interest when $n > p$. In our epidemiology example a sparse model has the interpretation of identifying a subset of potential risk factors that are most likely to be associated with the disease status. A natural thought given (1.1) is to take L to be the logistic deviance and take J to be the Elastic Net penalty. Indeed, there are several implementations available that perform the logistic regression using the logistic deviance loss and an Elastic Net penalty. Friedman et al. provide an R package GLMNET that performs Elastic Net penalized maximum likelihood estimation for generalized linear models [22]. Genkin et al. worked out a Bayesian formulation employing the Laplace prior and applied their model to text classification [24]. Wu et al. proposed minimizing a LASSO penalized logistic likelihood for genome wide association studies [57]. Finally, Liu et al. presented algorithms for both Elastic Net and concave Bridge penalized logistic likelihood models [36].

1.4 Model Misspecification

The shrinkage methods described above address regression and variable selection when the design matrix is ill-conditioned. As discussed above, the standard approach is to minimize a regularized negative log-likelihood function. Again the rationale for this approach is that under very general regularity conditions the maximum likelihood estimation procedure is asymptotically efficient. No other estimation procedure will have less variance as the number of observations increases. This efficiency comes at a cost, however, as maximum likelihood estimation is typically not robust to contaminants, or put another way, to model

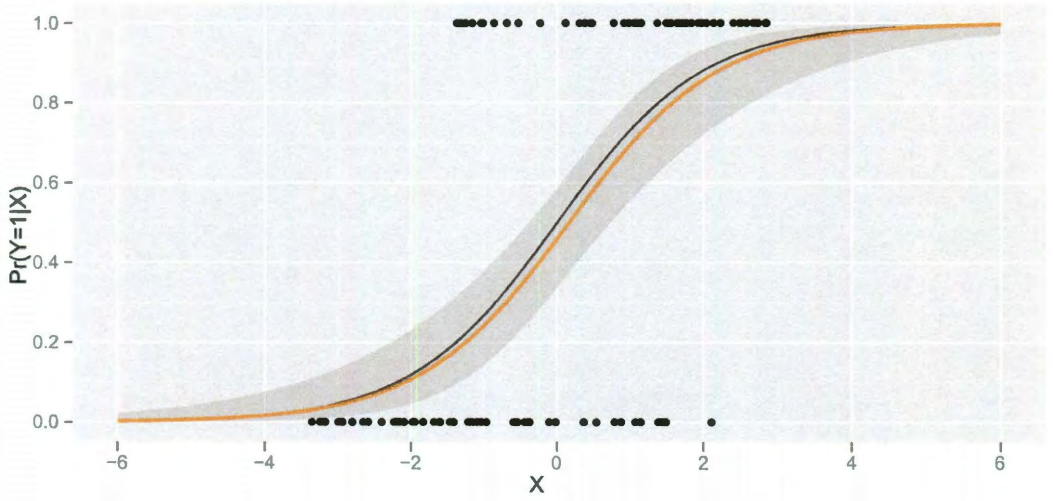


Figure 1.1: Univariate Logistic Regression: No outliers. Truth is in black, and the MLE is in orange with the grey region denoting a pointwise 95% confidence band.

misspecification.

To illustrate this claim consider the case where there is only one predictor. Figure 1.1 shows binary data generated according to a logistic model. The black line corresponds to the true generative distribution. The yellow line shows the MLE estimate of the distribution. Note the good agreement.

Now consider what happens when a few outliers are added. Figure 1.2 shows the resulting fits for the same data when 5 outliers are placed at extreme negative values for the covariate. We see that the yellow line corresponding to the MLE has been distorted, specifically it has "flattened." Recall that the slope of a univariate logistic function $F(\theta x)$ at the origin is $\theta/4$. Thus, the flatter the logistic curve is in the transition region, the smaller the corresponding regression coefficient. For this reason, the likelihood based logistic regression is said to suffer from "implosion" breakdown when outliers are added. If there are p covariates Croux et al. proved that it is always possible to add $2p$ outliers that will cause the magnitude of the logistic MLE $\|\hat{\boldsymbol{\theta}}\|_2$ to tend to zero [15]. Let Z_n denote the sample

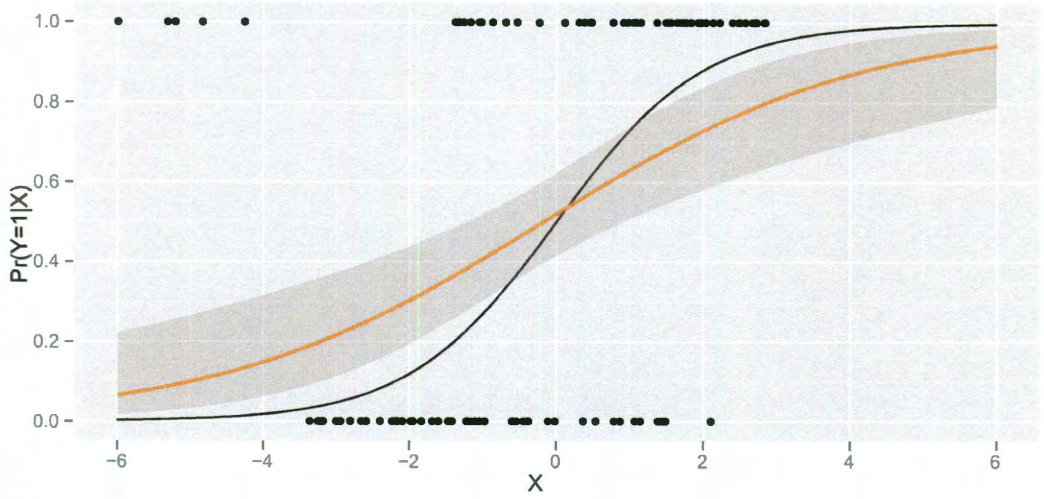


Figure 1.2: Univariate Logistic Regression: 5 outliers. Truth is in black, and the MLE is in orange with the grey region denoting a pointwise 95% confidence band.

$\{(X_i, Y_i)\}$ and $\hat{\boldsymbol{\theta}}^{(n)}$ correspond to the MLE given Z_n (2.3). Croux et al. define the breakdown point as the smallest fraction of contamination in the data that can cause the estimate $\hat{\boldsymbol{\theta}}^{(n)}$ to either grow to infinity (explode) or vanish (implode).

Definition 1. *The breakdown point of the estimator* $\hat{\boldsymbol{\theta}}^{(n)}$ *in (2.3) for a sample* Z_n *is given by* $\varepsilon^*(\hat{\boldsymbol{\theta}}^{(n)}; Z_n) = m^*/(n + m^*)$ *with* $m^* = \min(m^+, m^-)$,

$$m^+ = \min\left\{ m \in \mathbb{N}_0 : \sup_{\{(X_i,Y_i)\}_{i=n+1}^{n+m}} \|\hat{\theta}^{(n+m)}\|_2 = \infty \right\}$$

$$m^- = \min\left\{ m \in \mathbb{N}_0 : \inf_{\{(X_i,Y_i)\}_{i=n+1}^{n+m}} \|\hat{\theta}^{(n+m)}\|_2 = 0 \right\}.$$

Theorem 1 in [15] proves that the MLE never explodes as the result of adding outliers. So, m^* always equals m^-. We now restate Croux et al.'s result that implosion breakdown can occur with choice placement of additional observations.

Theorem 1 (Theorem 2 in [15]). *At any sample,* Z_n*, the breakdown point of the estimator*

$\hat{\theta}^{(n)}$ in (2.3) satisfies

$$\varepsilon^*(\hat{\theta}^{(n)}; Z_n) \leq \frac{2p}{n+2p}.$$

The example shown in Figure 1.2 demonstrates that breakdown can be the result of points of high leverage. Later in this thesis we will show that inliers or points of low leverage can also cause implosion breakdown. Moreover, scenarios will be explored where inliers and outliers both can cause serious trouble for variable selection.

Again while a great deal of work has been done for regularized negative log-likelihood loss functions relatively little has been explored with robust loss functions. The existing work centers around the Huber loss function. See Rosset and Zhu [42] as well as Owen [41]. Rosset and Zhu [42] as well as Wang et al.[55] discuss using a Huberized hinge loss for regularized robust classification. In these references, regularized robust estimation procedures are introduced, but compelling scenarios which justify their use are not explored.

The aim of this thesis is twofold; to extend regularization based classification and variable selection methods developed for the standard negative log-likelihood loss to losses that correspond to robust estimation using parametric models as well as to explore under what circumstances these robust variants are worth the extra computational trouble. To this end we propose a novel robust binary regression procedure based on minimizing the integrated square error (ISE) [47, 49], and to handle variable selection we consider an Elastic Net penalized minimum ISE loss.

The rest of this thesis will proceed as follows. In Chapter 2 we review minimum distance estimation with the ISE and derive the logistic regression version of it. In Chapter 3 we present our iterative algorithm for solving the L_2E optimization problem. In Chapter 4 we prove the global convergence of our algorithm. In Chapter 5 we present simulation results. In Chapter 6 we present real data analysis results. Chapter 7 concludes with a discussion and future directions of research.

CHAPTER 2

ROBUST BINARY LOGISTIC REGRESSION

We begin this chapter with a general discussion of robust estimation through the use of minimum ISE estimator or L_2E, and then discuss a variant aimed at fitting a partial parametric density. We then will derive the logistic L_2E loss and introduce the Elastic Net regularized version of it that will be used for all the examples and data sets where $n \ll p$. We then provide some intuition why the ISE criterion produces robust estimates before walking through an illustrative example comparing the logistic L_2E and the logistic MLE. This chapter concludes with a review of literature on robust logistic regression.

2.1 The Integrated Squared Error Loss

Let $f_{\boldsymbol{\theta}}$ be the density, indexed and completely specified by a parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ for some $p \in \mathbb{N}$, believed to be generating real valued data $Y_1, \ldots, Y_n$. Let f be the unknown true density generating the data. If we suspect outlying subgroups may be present within our data, we may wish to consider an alternative optimization problem to minimizing the negative log-likelihood. If we actually knew the true distribution, an intuitively good solution is the one that is "closest" to the true distribution. Consequently as an alternative to using the negative log-likelihood we consider the integrated square error (ISE) as the

loss function. Thus, we seek $\hat{\theta} \in \Theta$ that minimizes

$$\int [f_\theta(y) - f(y)]^2 \, dy. \tag{2.1}$$

Although finding such a θ is impossible since f is unknown, it is possible to find a θ that minimizes an unbiased estimate of the ISE. Expanding the integrand in (2.2), gives us

$$\int f_\theta(y)^2 \, dy - 2 \int f_\theta(y) f(y) \, dy + \int f(y)^2 \, dy.$$

The second integral is an expectation $E[f_\theta(Y)]$ where Y is a random variable drawn from a density f. This integral can be estimated from the data by the sample mean. The third integral does not depend on θ. With these observations in mind, we use the following loss function

$$L(\theta) = \int f_\theta(y)^2 \, dy - \frac{2}{n} \sum_{i=1}^{n} f_\theta(y_i)$$

and seek a $\hat{\theta}$ such that $L(\hat{\theta}) = \min_{\theta \in \Theta} L(\theta)$. The estimate $\hat{\theta}$ is called a L_2 estimate or L_2E by Scott [47].

Note that the minimization problem is a familiar one associated with bandwidth selection for histograms and more generally for kernel density estimators [48]. Applying a commonly used criterion in non-parametric density estimation to parametric estimation has the interesting consequence of trading off efficiency with robustness in the estimation procedure. In fact, previously Basu et al. have described a family of divergences which includes the L_2E as a special case and the MLE as a limiting case. The members of this family of divergences are indexed by a parameter that explicitly trades off efficiency for robustness [3]. They propose the following density power divergence between f and f_θ

$$d_\gamma(f, f_\theta) = \int \left\{ f_\theta^{1+\gamma}(z) - \left(1 + \frac{1}{\gamma}\right) f(z) f_\theta^\gamma(z) + \frac{1}{\gamma} f^{1+\gamma}(z) \right\} dz \quad ,$$

where $\gamma > 0$ is a tuning parameter which trades off robustness for efficiency. This loss includes the MLE as a special case since $d_\gamma(g, f)$ converges to the Kullback-Leibler divergence as $\gamma \to 0$ and minimizing the Kullback-Leibler divergence is equivalent to maximizing the likelihood function. The MLE is the most efficient but least robust member in this family of estimation procedures. When $\gamma = 1$ we recover the loss associated with the L_2E.

In [47, 46] Scott demonstrated that the L_2E has two benefits, the aforementioned robustness properties and computational tractability, all at a drop in asymptotic efficiency similar to that seen in comparing the mean and median as a location estimator.

2.1.1 Fitting partial densities

Scott provides examples where it is beneficial to find the minimum distance partial density estimate [46, 49]. The L_2E loss can be generalized to fit a parametric model to only the fraction, $w \in (0, 1]$, of the data that is described well by the parametric model:

$$\int \left[w f_{\boldsymbol{\theta}}(y) - f(y) \right]^2 dy. \tag{2.2}$$

We define the loss measuring the lack of fit of between the data and a partial density:

$$L(\boldsymbol{\theta}, w) = \int w^2 f_{\boldsymbol{\theta}}(y)^2 \, dy - \frac{2w}{n} \sum_{i=1}^{n} f_{\boldsymbol{\theta}}(y_i).$$

The L_2E then becomes the pair $(\hat{\boldsymbol{\theta}}, \hat{w})$ such that

$$L(\hat{\boldsymbol{\theta}}, \hat{w}) = \min_{\boldsymbol{\theta} \in \Theta, w \in (0,1]} L(\boldsymbol{\theta}, w).$$

Note that w need not be specified *a priori* but may be a parameter to be optimized over and estimated from the data. Note that $L(\boldsymbol{\theta}, w)$ may have multiple local minima.

2.2 Logistic L_2E

We now adapt the L_2E method to logistic regression. Let $Y_1, \ldots, Y_n$ denote n binary response random variables, y_i denote the ith observed value and $\mathbf{x}_i$ denote its associated p-dimensional covariate. Again the standard approach to estimating the conditional probability is to fit a logistic model by maximum likelihood estimation [38]. Let $F(t) = (1 + \exp(-t))^{-1}$ denote the logistic function. In logistic regression the mean response is modeled as a logistic function of a linear combination of the covariates

$$P_{\boldsymbol{\theta}}(Y_i = y|\mathbf{x}_i) = \begin{cases} 1 - F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}) & y = 0 \\ F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}) & y = 1 \end{cases}, \tag{2.3}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$. We adopt the convention that the first element of $\boldsymbol{\theta}$, denoted θ_0, is an intercept term. Thus, we set the first element of the every covariate vector to be 1. Let $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ be the matrix of covariates whose first column consists of ones. Recall $\mathbf{x}_i$ denotes the transpose of the ith row of $\mathbf{X}$. The loss function we minimize is the average L_2E loss which is given by

$$L(\boldsymbol{\theta}, w) = \frac{1}{n}\sum_{i=1}^{n}\left\{ w^2 \sum_{y\in\{0,1\}} P_{\boldsymbol{\theta}}(Y_i = y|\mathbf{x}_i)^2 - 2wP_{\boldsymbol{\theta}}(Y_i = y_i|\mathbf{x}_i) \right\}. \tag{2.4}$$

Finally, to deal with high dimensional covariates that may be correlated we penalize our loss with the Elastic Net penalty. Thus, we seek a $\hat{\boldsymbol{\theta}}(\lambda)$ such that

$$L(\hat{\boldsymbol{\theta}}(\lambda), w) + \lambda J(\hat{\boldsymbol{\theta}}(\lambda)) = \min_{\boldsymbol{\theta}} \left\{ L(\boldsymbol{\theta}, w) + \lambda J(\boldsymbol{\theta}) \right\},$$

where $J(\boldsymbol{\theta}) = \alpha \| \boldsymbol{\theta} \|_1 + \frac{(1-\alpha)}{2} \| \boldsymbol{\theta} \|_2^2$ and $\alpha \in [0, 1]$ trades off the relative strengths of the LASSO and ridge penalties.

2.3 Intuition

Understanding why the ISE imparts robustness requires understanding why the MLE can be very sensitive to outliers. Revisit Figure 1.2. Note that the distortion in the MLE is due to the MLE striving to find a logistic distribution that has positive probability that can account for the five outliers present at large negative values. Because the parametric model is not flexible, however, this comes at a cost of choosing a model that substantially increases the probability of observing a 1 when the covariate values are in the interval $(-4, -2)$ for example as well as substantially increasing the probability of observing a 0 when the covariate values are in the interval $(2, 3)$. The MLE is overly aggressive about finding a parametric model that puts probability mass where the outliers are observed but does so at the expense of poorly modeling regions where no data is observed.

The L_2E in contrast is more conservative and seeks a parametric model that balances placing probability mass in accord where data are observed against not putting too much probability mass where data is not observed.

2.4 An Illustrative Example

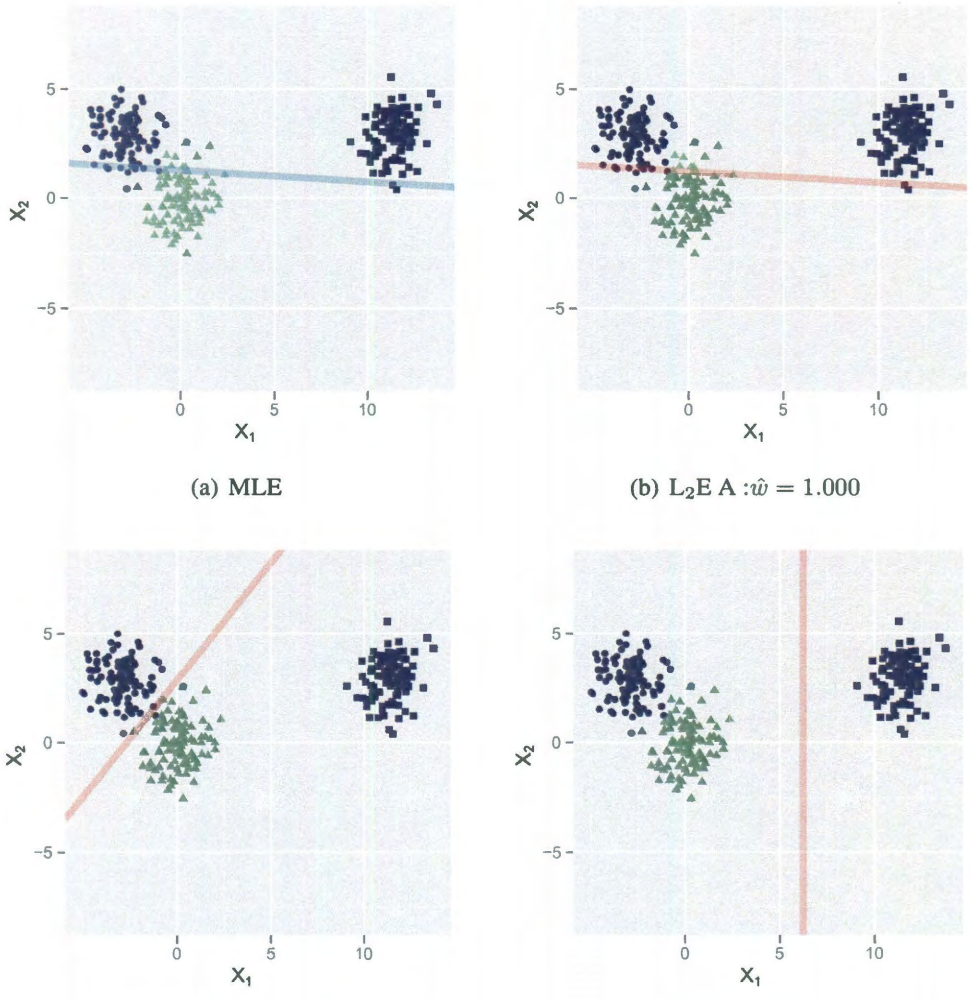
To better understand the behavior of the L_2E and illustrate how drastically the MLE and L_2E solutions can differ, consider the following simple example where $n > p$. We observe 300 binary outcomes. Let $\mathbf{x}_i \in \mathbb{R}^3$ denote the ith observed covariate. The first element x_{i0} is 1 for every i. For $i = 1, \ldots, 100$, x_{i1} are iid $N(-3, 1)$ and x_{i2} are iid $N(3, 1)$. For $i = 101, \ldots, 200$, x_{i1} and x_{i2} are iid $N(0, 1)$. For $i = 201, \ldots, 300$, x_{i1} are iid $N(11.5, 1)$, and x_{i2} are iid $N(3, 1)$. The binary outcomes were assigned as follows: $y_i = 1$ for $i = 101, \ldots, 200$ and 0 otherwise.

This construction results in three equally sized clusters, two which carry the 0 label and one which carries the 1 label. We compare the solutions to the L_2E optimization problem

and the maximum likelihood estimation for this data set. Figure 2.1(a) shows the median line for the MLE fit. The median line of the fitted logistic surface are computed with the **nlminb** iterative solver in R. Three lines from different initial values for θ are shown in Figures 2.1(b), 2.1(c), 2.1(d). The iterative solver for the L_2E loss converged to three distinct local minima. We denote the resulting fits L_2E A, B, and C. L_2E A agrees with the maximum likelihood estimate. So, the MLE solution is a local minima of the ISE loss, in fact in this example the L_2E A solution is the global minimum. The L_2E loss for this fit is -0.889, and $\hat{w} = 1$. The fits corresponding to the other two local minima, L_2E B and C, had corresponding loss values of -0.433 and -0.391, and for both L_2E B and C, $\hat{w} \sim 2/3$.

The parameter w gives the extra degree of freedom needed to fit the parametric model to some of the data but not necessarily all of it. The fitted value of w represents the fraction of data that follows the parametric model. Again in practice we would not know ahead of time that there are two subclasses within the zero subclass. As a diagnostic for outliers we can do jitter plots of the predicted probabilities against the observed labels. Figures 2.2(a), 2.2(b), and 2.2(c) show such plots for the given example. Samples with fitted probabilities that are most inconsistent with the observed values should be investigated as possible outliers.

This simple example illustrates the strengths of the L_2E. The underlying mixture is fit *without* pre-specifying a model for the portion of data that does not appear to come from the parametric model. The unfortunate reality is that it is unlikely that there is good *a priori* knowledge of how much contaminant exists in the data and what model would suitably describe the contamination. If this were known, robust regression would be unnecessary. In contrast, the MLE can be a bit Procrustean. It forces *all* the data to come from the parametric model. We see that the MLE chooses a separating hyperplane that indeed separates the 0's from the 1's. But it fails to distinguish that there are two ways in covariate space to be a 0 as illustrated in Figure 2.2(b).



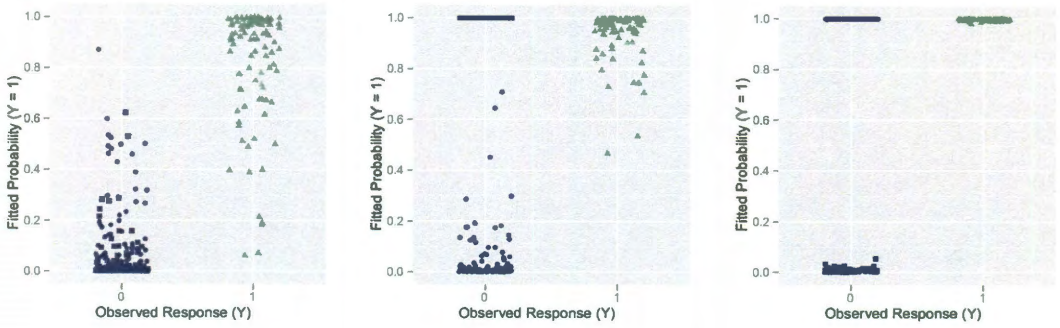
(a) MLE

(b) L_2E A :$\hat{w} = 1.000$

(c) L_2E B: $\hat{w} = 0.666$

(d) L_2E C: $\hat{w} = 0.668$

Figure 2.1: An example where $n = 300$ and $p = 2$ to contrast the MLE and L_2E solutions. Responses are binary. Blue circles and squares denotes zeros; green triangles denotes ones. The decision boundary is the median line of the resulting logistic fit. The L_2E method finds three local minima: (b), (c) and (d). Note that (b) finds a model that fits all the data and not surprisingly reproduces the MLE. On the other hand (c) and (d) find models that fit only two thirds of the data and match what the MLE would produce if only those two thirds of the data were used.



(a) L_2E A (b) L_2E B (c) L_2E C

Figure 2.2: Jitter plots of the fitted probabilities of being 1 against their observed responses: Solutions L_2E B and C identify the presence of outliers. There is a discrepancy between the fitted and observed responses in (b) and (c). Outliers are indistinguishable from the rest of the data in L_2E A, (a). The corresponding jitter plot for the MLE solution is not shown since it is identical to the one corresponding to L_2E A.

2.4.1 Related Work

Kim and Scott applied the L_2E to classification using kernel density estimates [29, 30]. Their problem formulation resulted in a quadratic optimization problem similar to the one performed in training a support vector machine. They obtained a classifier that had a sparse representation but not in the original predictors as our method finds. Moreover, their focus was not in answering questions about the robustness imparted by the ISE criterion. Several prior works have proposed robust methods of logistic regression. A broad class of solutions consists of downweighting the contribution of outlying points to the estimating equations. Downweighting can be based on extreme values in covariate space [32, 9] or on extreme predicted probabilities [14, 9, 4]. These estimators have estimating equations that have the following form

$$\mathbf{0} = \sum_{i=1}^{n} w_i \mathbf{x}_i \{Y_i - P_{\boldsymbol{\theta}}(Y_i = 1 | X_i = \mathbf{x}_i) - c(\mathbf{x}_i, \boldsymbol{\theta})\}.$$

The w_i are weights associated with each observation; they may be a function of the co-

variates, response, or both. The term $c(\mathbf{x}_i, \boldsymbol{\theta})$ is included to control the bias of the estimates. If $w_i = 1$ and $c(\mathbf{x}_i, \boldsymbol{\theta}) = 0$ then we recover the standard logistic regression coefficients. Taking $w_i = w(\mathbf{x}_i, \mathbf{x}_i^\mathsf{T}\boldsymbol{\theta})$ and $c(\mathbf{x}_i, \mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}) = 0$ gives us the so-called "Mallows" class which can be used to make estimates robust against extreme values in the covariate space. Taking $w_i = w(\mathbf{x}_i, \mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}, y_i)$ gives us the so-called "Schweppe" class which gives the most generality and can be used to make estimates robust against extreme values in both the covariate space and errors in the recorded response.

Interestingly the L_2E estimating equation looks quite similar

$$0 = \sum_{i=1}^{n} \gamma_i \mathbf{x}_i \{Y_i - w P_{\boldsymbol{\theta}}(Y_i = 1 | X_i = \mathbf{x}_i) - \left(\frac{1-w}{2}\right)\}$$

where

$$\gamma_i = P_{\boldsymbol{\theta}}(Y_i = 1 | X_i = \mathbf{x}_i) P_{\boldsymbol{\theta}}(Y_i = 0 | X_i = \mathbf{x}_i).$$

Furthermore, when $w = 1$, the L_2E estimator reduces to a "Mallows" class estimator.

The work by Bondell in [6] is similar to ours in that he considered fitting parameters by minimizing a weighted Cramér-von Mises distance. Our work diverges from his in that we do not explicitly use estimates of the underlying distribution.

The fundamental difference between the approach proposed here and prior work is the application of regularization to handle high dimensional data and perform variable selection in the presence of model misspecification. The choice of the ISE as a loss is motivated by its lack of a tuning parameter and by virtue that there is a simple quadratic approximation which facilitates its computation as will be seen in Chapter 3.

CHAPTER 3

ALGORITHMS FOR ESTIMATION AND VARIABLE SELECTION WITH THE LOGISTIC L_2E

General solvers like **nlminb** are adequate as off the shelf solutions for logistic L_2E regression in low dimensions such as the illustration given in Section 2.4. For larger problems and those which are regularized by a non-differentiable penalty, we develop a coordinate descent algorithm [21, 58]. The logistic L_2E loss is not convex. In fact, the Hessian of the L_2E loss is non-negative definite for some values of θ, which rules out the use of unconstrained second order optimization methods like the standard Newton's Method. Instead, we minimize the L_2E loss with a majorization-minimization (MM) algorithm [27].

The rest of Chapter 3 is organized as follows. We first review the MM algorithm. We then present a convex quadratic majorization of the logistic L_2E loss derived in Chapter 2. Finally we conclude Chapter 3 with two iterative algorithms for carrying out the MM algorithm. The former is suitable for the classical case when $n > p$ and no regularization is applied. The latter is suitable when $n \ll p$ and the elastic net penalty is applied. We rely on coordinate descent to minimize each quadratic majorization.

Algorithm 1 MAJORIZATION-MINIMIZATION

$\mathbf{x}_0 \leftarrow$ initial guess
$h_0 \leftarrow$ majorization of g at $\mathbf{x}_0$.
$k \leftarrow 0$
repeat
  $\mathbf{x}_{k+1} \leftarrow \underset{\mathbf{x}}{\text{argmin}}\ h_k(\mathbf{x})$
  $h_{k+1} \leftarrow$ majorization of g at $\mathbf{x}_{k+1}$
  $k \leftarrow k + 1$
until convergence
return $\mathbf{x}_{k+1}$

3.1 Majorization-Minimization

The strategy behind majorization-minimization is to minimize a surrogate function, the majorization, instead of the original objective function. The surrogate is chosen with two goals in mind. First, an argument that decreases the surrogate should decrease the objective function. Second, the surrogate should be easier to minimize than the objective function. We give the formal definition of majorization.

Definition 2. *Suppose g and h are real-valued functions on $\mathbb{R}^p$. We say that h* majorizes *g at $\mathbf{x}$ if $h(\mathbf{u}) \geq g(\mathbf{u})$ for all $\mathbf{u}$ and $h(\mathbf{x}) = g(\mathbf{x})$.*

In words, the surface h lies above the surface g and is tangent to g at $\mathbf{x}$. Algorithm 1 shows a simple iterative algorithm for finding the minimum of f provided we can find a majorization for f at every point in $\mathbb{R}^p$. It is easy to see that Algorithm 1 always takes non-increasing steps with respect to g for the following reason. Consider the iteration starting at $\mathbf{x}_k$. Since $\mathbf{x}_{k+1}$ minimizes h_k, we have

$$g(\mathbf{x}_k) = h_k(\mathbf{x}_k) \geq h_k(\mathbf{x}_{k+1}) \geq g(\mathbf{x}_{k+1}).$$

By using Algorithm 1, we can convert a hard optimization problem (e.g. non-convex, non-differentiable) into a series of simpler ones (e.g. smooth convex), each of which is easier to minimize than the original.

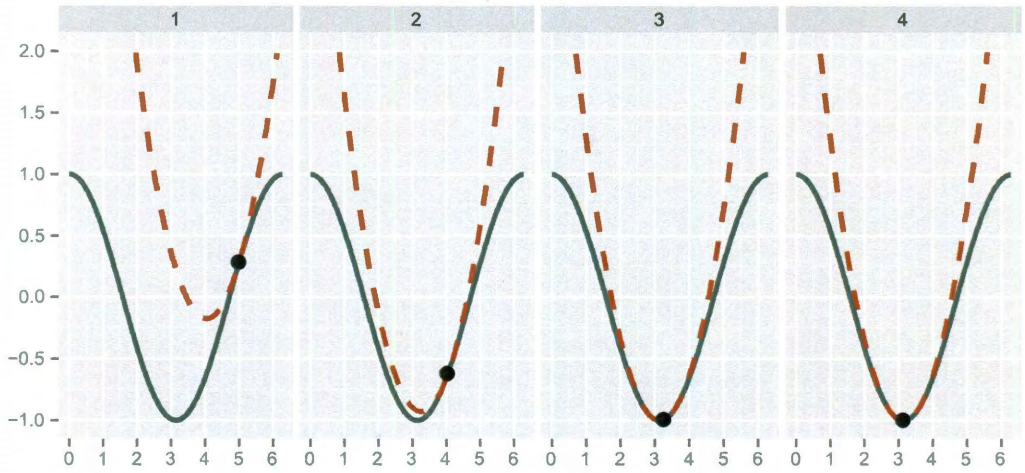


Figure 3.1: Four iterations of an MM algorithm minimizing the cosine function. The majorization is drawn in orange dashed lines

Figure 3.1 shows an example of an MM algorithm used to find local minima of the cosine function. Note that the cosine function has an exact second order Taylor expansion at an arbitrary point x' in its domain.

$$\cos(x) = \cos(x') - \sin(x')(x - x') - \cos(x^*)(x - x')^2$$

for some x^* between x and x'. Furthermore $-\cos(x^*)$ is bounded above by 1. Therefore, a simple majorization for $\cos(x)$ at x' is given by the convex quadratic function

$$\cos(x') - \sin(x')(x - x') + (x - x')^2.$$

This approach applies in general to functions with continuous second derivative and bounded curvature [5]. In the next section we use the fact that the logistic L_2E loss has bounded curvature to derive a convex quadratic majorization in a like manner.

3.2 Majorizing the L_2E loss

Recall that the minimization problem at hand is

$$L(\hat{\boldsymbol{\theta}}, \hat{w}) = \min_{\boldsymbol{\theta}\in\mathbb{R}^{p+1}, w\in(0,1]} L(\boldsymbol{\theta}, w).$$

For the rest of this paper, we do not jointly optimize over w and $\boldsymbol{\theta}$. Instead we fix w and optimize over $\boldsymbol{\theta}$. In simulation experiments in high dimensions with $n \ll p$, preliminary attempts to jointly optimize over $\boldsymbol{\theta}$ and w led to non-convergent behavior in the iterated estimates of w. Instead we perform the following optimization over a grid of w values. For fixed $w \in (0, 1]$, let $L_w(\boldsymbol{\theta})$ denote $L(\boldsymbol{\theta}, w)$ as a function of $\boldsymbol{\theta}$ only. We then seek a $\hat{\boldsymbol{\theta}}$ such that

$$L_w(\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}\in\mathbb{R}^{p+1}} L(\boldsymbol{\theta}, w). \tag{3.1}$$

We use the following convex quadratic majorization of $L_w(\boldsymbol{\theta})$ to find $\hat{\boldsymbol{\theta}}$. A proof of Theorem 2 is given in Section 3.7.

Theorem 2. *Suppose $w \in (0, 1]$, $\mathbf{X}$ is an n by $p+1$ matrix of covariates, and $\mathbf{y}$ is a vector of n binary responses, then the following function, $L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$, majorizes $L_w(\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$:*

$$L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = L_w(\tilde{\boldsymbol{\theta}}) + 2\frac{w}{n}\mathbf{z}_{\tilde{\boldsymbol{\theta}}}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \frac{w}{n}\eta\|\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\|_2^2,$$

where $\mathbf{z}_{\tilde{\boldsymbol{\theta}}} = \mathbf{G}(w\mathbf{q} - \boldsymbol{\upsilon})$ is a working response and is defined by the following quantities:

$$\boldsymbol{\upsilon} = 2\mathbf{y} - \mathbf{1}, \quad \mathbf{q} = 2\mathbf{p} - \mathbf{1}, \quad \mathbf{p} = F(\mathbf{x}_i^{\mathsf{T}}\tilde{\boldsymbol{\theta}}),$$

and $\mathbf{G} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with $g_{ii} = p_i(1-p_i)$, and η is a design parameter such that

$$\eta \geq \frac{1}{4}\max_{a\in[-1,1]}\left\{\frac{3}{2}wa^4 - a^3 - 2wa^2 + a + \frac{w}{2}\right\}.$$

In the next section we will see that the parameter η^{-1} controls the step size of our iterative solver. Consequently, in practice we set η to its lower bound to take the largest steps possible to speed up convergence. Consider the polynomial

$$f(a) = \frac{3}{2}wa^4 - a^3 - 2wa^2 + a + \frac{w}{2}.$$

When $w = 1$, f' has a root at 1, and it is straightforward to show that the lower bound of η is attained at the second largest root of f', which is $(-3+\sqrt{33})/12$. In fact the maximizing arguments of f on $[-1, 1]$ will always be the second largest root of f' for $w \in (0, 1]$. We use this fact to numerical determine optimal values of η as a function of w.

3.3 Solving by Iterative Least Squares

In this and the next two sections we use the convention that $\boldsymbol{\theta}$ denotes the last p coordinates of the parameter vector $(\theta_0, \boldsymbol{\theta}) \in \mathbb{R} \times \mathbb{R}^p$; the first column of the data matrix $\mathbf{X}$ is no longer $\mathbf{1}$. We make this change for clarity of the derivation of the coordinate descent rules. Note that by completing the square we can compactly express the majorization $L_w(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ as the sum of terms that depend on $\boldsymbol{\theta}$ and those that do not.

$$L_w(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \frac{w}{n}\eta\|\zeta(\tilde{\boldsymbol{\theta}}) - \theta_0\mathbf{1} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + L_w(\tilde{\boldsymbol{\theta}}) - \frac{w}{n}\frac{1}{\eta}\|z_{\tilde{\boldsymbol{\theta}}}\|_2^2,$$

where $\zeta(\tilde{\boldsymbol{\theta}}) = \tilde{\theta}_0\mathbf{1} + \mathbf{X}\tilde{\boldsymbol{\theta}} - \eta^{-1}z_{\tilde{\boldsymbol{\theta}}}$. We will assume that the columns of $\mathbf{X}$ are centered. Setting the gradient of $L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$ with respect to θ_0 and $\boldsymbol{\theta}$ to zero gives the normal equations

$$\begin{aligned} \theta_0 &= \tilde{\theta}_0 - \eta^{-1}\overline{z_{\tilde{\boldsymbol{\theta}}}} \\ \mathbf{X}^\mathsf{T}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) &= -\frac{1}{\eta}\mathbf{X}^\mathsf{T}\mathbf{z}_{\tilde{\boldsymbol{\theta}}}, \end{aligned} \tag{3.2}$$

where $\overline{z_{\tilde{\boldsymbol{\theta}}}} = n^{-1}\mathbf{1}^\mathsf{T}\mathbf{z}_{\tilde{\boldsymbol{\theta}}}$.

Algorithm 2 ITERATIVE L_2E SOLVER

$(\theta_0, \boldsymbol{\theta}^\mathsf{T}) \leftarrow$ initial guess
$\boldsymbol{\upsilon} \leftarrow 2\mathbf{y} - \mathbf{1}$
repeat
  $\mathbf{p} \leftarrow F(\theta_0\mathbf{1} + \mathbf{X}\boldsymbol{\theta})$
  $\mathbf{q} \leftarrow 2\mathbf{p} - \mathbf{1}$
  $\mathbf{G} \leftarrow \text{diag}(\mathbf{p} * (\mathbf{1} - \mathbf{p}))$
  $\mathbf{z} \leftarrow \mathbf{G}(w\mathbf{q} - \boldsymbol{\upsilon})$
  $\theta_0 \leftarrow \theta_0 - \eta^{-1}\bar{z}$
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \frac{1}{\eta}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\mathsf{T}\mathbf{z}$
until convergence
return $(\theta_0, \boldsymbol{\theta}^\mathsf{T})$

If we compute the singular value decomposition (SVD) of $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\mathsf{T}$ where $\mathbf{U} \in \mathbb{R}^{n\times k}$ and $\mathbf{V} \in \mathbb{R}^{p\times k}$ have columns that are orthonormal, k is the rank of $\mathbf{X}$, and $\mathbf{D} \in \mathbb{R}^{k\times k}$ is the diagonal matrix of singular values, then we can find a $\boldsymbol{\theta}$ that minimizes the majorization of the L_2E loss at $\tilde{\boldsymbol{\theta}}$ with the following update rule.

$$\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} - \frac{1}{\eta}\mathbf{V}\mathbf{D}^{-1}\mathbf{U}^\mathsf{T}\mathbf{z}_{\tilde{\boldsymbol{\theta}}}.$$

Note that the step size is inversely proportional to η. If $\mathbf{X}^\mathsf{T}\mathbf{X}$ is full rank, as is likely when $n > p$, then the $\boldsymbol{\theta}$ that solves (3.2) is unique. Algorithm 2 provides pseudocode for update rule for calculating the $(m+1)$th parameter vector, $\boldsymbol{\theta}_{m+1}$, from the mth one, $\boldsymbol{\theta}_m$.

3.4 Solving by Coordinate Descent

Alternatively we can minimize the majorizer $L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$ by coordinate descent, i.e. update each component of the parameter vector one by one. We consider this because the coordinate descent solver can be easily generalized to handle an ℓ_1 and Elastic Net penalty on $\boldsymbol{\theta}$.

Coordinate descent is a special case of block relaxation optimization where in a round robin fashion we optimize the objective function with respect to each coordinate at a time

while holding all other coordinates fixed. Formally at the kth iteration for the ith coordinate we solve

$$x_i^{(k)} \in \arg\min_u f(x_1^{(k)}, \ldots, x_{i-1}^{(k)}, u, x_{i+1}^{(k-1)}, \ldots, x_p^{(k-1)}).$$

We present the coordinate descent algorithm for minimizing the Elastic Net regularized logistic L_2E regression problem. Note that the majorization given in Theorem 2 can be adapted for regularization. It follows immediately that $(1/2)L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) + \lambda J(\boldsymbol{\theta})$ majorizes $(1/2)L_w(\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta})$ for a penalty function $J : \mathbb{R}^p \to \mathbb{R}$. In particular, consider the penalized majorizer for the L_2E loss regularized by the Elastic Net penalty,

$$\frac{1}{2}L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) + \lambda \left(\alpha \|\boldsymbol{\theta}\|_1 + \frac{(1-\alpha)}{2} \|\boldsymbol{\theta}\|_2^2 \right).$$

Since θ_0 is not penalized the update for it is unchanged; it is still the mean of the working response $\mathbf{z}_{\tilde{\boldsymbol{\theta}}}$. The kth coordinate update during the mth round of iteration is well defined, i.e. there exists a unique minimizer in (3.4), and is given by

$$\theta_k^m = \arg\min_{\theta_k} \left\{ \frac{1}{2}\frac{w}{n}\eta \|\tilde{\boldsymbol{\zeta}} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \left(\alpha \|\boldsymbol{\theta}\|_1 + \frac{(1-\alpha)}{2} \|\boldsymbol{\theta}\|_2^2 \right) \right\},$$

where $\boldsymbol{\theta} = (\theta_1^m, \ldots, \theta_{k-1}^m, \theta_k, \theta_{k+1}^{m-1}, \ldots \theta_p^{m-1})$ and $\tilde{\boldsymbol{\zeta}} = \mathbf{X}\tilde{\boldsymbol{\theta}} - \eta^{-1}(\mathbf{z}_{\tilde{\boldsymbol{\theta}}} - \overline{z_{\tilde{\boldsymbol{\theta}}}}\mathbf{1})$. With a little convex analysis it is straightforward to show that

$$\theta_k^m = \frac{S(\frac{w}{n}\eta \sum_{i=1}^n x_{ik}(\tilde{\boldsymbol{\zeta}}_i - (\sum_{j=1}^{k-1} x_{ij}\theta_j^m + \sum_{j=k+1}^p x_{ij}\theta_j^{m-1})), \lambda\alpha)}{\frac{w}{n}\eta \sum_{i=1}^n x_{ik}^2 + \lambda(1-\alpha)}$$

where S is the soft threshold function

$$S(a, \lambda) = \text{sign}(a) \max(|a| - \lambda, 0).$$

The update is derived in Section 3.8.

Algorithm 3 ITERATIVE L_2E SOLVER

$(\theta_0, \boldsymbol{\theta}^\mathsf{T}) \leftarrow$ initial guess
$\boldsymbol{\upsilon} \leftarrow 2\mathbf{y} - \mathbf{1}$
repeat
  $\mathbf{p} \leftarrow F(\theta_0\mathbf{1} + \mathbf{X}\boldsymbol{\theta})$
  $\mathbf{q} \leftarrow 2\mathbf{p} - \mathbf{1}$
  $\mathbf{G} \leftarrow \mathrm{diag}(\mathbf{p} * (\mathbf{1} - \mathbf{p}))$
  $\mathbf{z} \leftarrow \mathbf{G}(w\mathbf{q} - \boldsymbol{\upsilon})$
  $\boldsymbol{\zeta} \leftarrow \mathbf{X}\boldsymbol{\theta} - \frac{1}{\eta}(\mathbf{z} - \overline{z}\mathbf{1})$
  $\theta_0 \leftarrow \theta_0 - \eta^{-1}\overline{z}$
  repeat
    for $k = 1..p$ **do**
      $\theta_k \leftarrow \frac{S(\frac{w}{n}\eta\sum_{i=1}^n x_{ik}(\zeta_i - \sum_{j\neq k} x_{ij}\theta_j), \lambda\alpha)}{\frac{w}{n}\eta\sum_{i=1}^n x_{ik}^2 + \lambda(1-\alpha)}$
    end for
  until convergence
until convergence
return $(\theta_0, \boldsymbol{\theta}^\mathsf{T})$

Algorithm 3 gives pseudocode for the resulting iterative solver. Note that $*$ denotes the Hadamard element-wise product, $F(\theta_0\mathbf{1} + \mathbf{X}\boldsymbol{\theta})$ is evaluated component-wise, and $\mathrm{diag}(\cdot)$ takes a vector of length n and puts it along the diagonal of an n-by-n diagonal matrix.

3.5 Warm Starts and Calculating Regularization Paths

Later we will need to compare the regression coefficients obtained at many values of the penalty parameter λ to perform model selection. We can rapidly calculate regression coefficients for a decreasing sequence of values of λ through warm starts. The idea is as follows. Suppose we calculate $\boldsymbol{\theta}(\lambda_1)$ the regression coefficients when the penalty is λ_1. If we wish to calculate $\boldsymbol{\theta}(\lambda_2)$ where λ_2 is a little bit smaller than λ_1 we should use $\boldsymbol{\theta}(\lambda_1)$ as the initial estimate for $\boldsymbol{\theta}(\lambda_2)$ in our coordinate descent algorithm. When λ_1 and λ_2 are close, the corresponding optimization problems are close. If the differences between our sequence of penalty parameters is small enough then each the solution of the preceding optimization problem will be close to the previous solution and the iterative algorithm should take less

time to find the optimum of the current problem. We proceed from largest penalty parameters to smallest because the larger the penalty is the easier the optimization problems is to solve.

We know that if λ is sufficiently large enough that only the intercept term θ_0 will come into the model. Let $p = F(\theta_0)$. In this case we have that

$$\begin{aligned} 0 &= \overline{z}_{\boldsymbol{\theta}} \\ 0 &= n^{-1}\mathbf{1}^{\mathsf{T}}\mathbf{G}(w\mathbf{q} - \boldsymbol{\upsilon}) \\ 0 &= n^{-1}p(1-p)\mathbf{1}^{\mathsf{T}}(w\mathbf{q} - \boldsymbol{\upsilon}) \\ w(2p-1) &= 2\overline{y} - 1 \end{aligned}$$

$$\theta_0 = \log\left(\frac{\overline{y} - \frac{1}{2}(1-w)}{\frac{1}{2}(1+w) - \overline{y}}\right) \tag{3.3}$$

Note that we must have that $\frac{1}{2}(1-w) < \overline{y} < \frac{1}{2}(1+w)$, or equivalently that $w > \max(2\overline{y} - 1, 1 - 2\overline{y})$. This condition is always met when $w = 1$ so long as there is at least one of each kind of response. The smallest λ^* such that all regression coefficients are shrunk to zero is given by

$$\begin{aligned} \lambda^* &= \frac{w}{n\alpha} \max_{j=1,\ldots,p} |\mathbf{x}_j^{\mathsf{T}}\mathbf{z}_{\theta_0}| \\ &= 2\frac{w}{n\alpha} F(\theta_0)(1 - F(\theta_0)) \max_{j=1,\ldots,p} |\mathbf{x}_j^{\mathsf{T}}\mathbf{y}|. \end{aligned} \tag{3.4}$$

We first compute θ_0 using (3.3) with $\lambda_{\text{max}} = \lambda^*$. We then set $\lambda_{\text{min}} = \epsilon\lambda_{\text{max}}$ for $\epsilon \ll 1$ and compute a grid of evenly spaced intermediate λ values equally spaced on the log scale between λ_{max} and λ_{min}. In practice, we have found the choice of $\epsilon = 0.05$ to be useful. In general, we are not interested in making λ so small as to include all variables.

3.6 Degrees of Freedom and Variable Selection

Once we have a set models computed at different regularization parameter values, we select the model that is optimal with respect to some criterion; often the criterion is prediction risk and the model with the least risk is selected. Two well known criteria are AIC [1] and BIC [45] both of which depend on the degrees of freedom in the model. Let $\ell(\boldsymbol{\theta})$ denote the likelihood then

$$\begin{aligned} \text{AIC}(\hat{\boldsymbol{\theta}}) &= -\frac{2}{n}\log(\ell(\hat{\boldsymbol{\theta}})) + \frac{2}{n}\text{df}(\hat{\boldsymbol{\theta}}) \\ \text{BIC}(\hat{\boldsymbol{\theta}}) &= -\frac{2}{n}\log(\ell(\hat{\boldsymbol{\theta}})) + \frac{\log(n)}{n}\text{df}(\hat{\boldsymbol{\theta}}) \end{aligned}$$

where $\text{df}(\hat{\boldsymbol{\theta}})$ denotes the degrees of freedom of the model at $\hat{\boldsymbol{\theta}}$.

Zou et al. [63] proved that an unbiased estimate of the degrees of freedom for the LASSO penalized least squares problem is given by the cardinality of the active set of variables (Theorem 1 in [63]). Note that the Elastic Net penalized Least Squares problem can be expressed as a LASSO penalized Least Squares problem.

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda\alpha\|\boldsymbol{\theta}\|_1 + \frac{\lambda(1-\alpha)}{2} = \frac{1}{2}\left\|\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} - \begin{bmatrix}\mathbf{X}\\ \sqrt{\lambda(1-\alpha)}\mathbf{I}\end{bmatrix}\boldsymbol{\theta}\right\|_2^2 + \lambda\alpha\|\boldsymbol{\theta}\|_1.$$

Thus, a simple modification of the proof of Theorem 1 in [63] gives an unbiased estimate for the degrees of freedom in an Elastic Net penalized model. Per [63] we denote by $\mathcal{B}_{\lambda,\alpha}$ the set of indices of the non-zero regression coefficients under penalty parameters λ and α. If we use the exact same arguments with the one change being to use the projection matrix onto the space $\mathbf{X}_{\mathcal{B}_{\lambda,\alpha}}$ to account for the contribution from the ridge penalty in the proof of

Theorem 1 to

$$H_{\lambda,\alpha} = \mathbf{X}_{\mathcal{B}_{\lambda,\alpha}} \left(\begin{bmatrix} \mathbf{X}_{\mathcal{B}_{\lambda,\alpha}} \\ \sqrt{\lambda(1-\alpha)}\mathbf{I} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} \mathbf{X}_{\mathcal{B}_{\lambda,\alpha}} \\ \sqrt{\lambda(1-\alpha)}\mathbf{I} \end{bmatrix} \right)^{-1} \mathbf{X}_{\mathcal{B}_{\lambda,\alpha}}^{\mathsf{T}}$$
$$= \mathbf{X}_{\mathcal{B}_{\lambda,\alpha}} \left(\mathbf{X}_{\mathcal{B}_{\lambda,\alpha}}^{\mathsf{T}} \mathbf{X}_{\mathcal{B}_{\lambda,\alpha}} + \lambda(1-\alpha)\mathbf{I} \right)^{-1} \mathbf{X}_{\mathcal{B}_{\lambda,\alpha}}^{\mathsf{T}},$$

then we obtain the following unbiased estimate of the degrees of freedom for the Elastic Net

$$\mathrm{df}(\hat{\boldsymbol{\theta}}) = \mathrm{tr}(H_{\lambda,\alpha}).$$

Note that when $\alpha = 1$ we recover the estimated degrees of freedom for the LASSO, $|\mathcal{B}_{\lambda,1}|$ and when $\alpha = 0$ we recover the degrees of freedom for ridge regression, $\mathrm{tr}(\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}})$.

Finally to perform model selection we do the following. We compute $(\hat{\theta}_0(\lambda), \hat{\boldsymbol{\theta}}(\lambda))$ for a sequence of λ's from $\lambda_{\max}$ down to $\lambda_{\min}$. Next we refit the model using the reduced variable set $\mathcal{B}_{\lambda,\alpha}$ and refit using logistic L_2E with $\alpha = 0$. This produces less biased estimates. We are adopting the same strategy as LARS-OLS in [17]. Our framework, however, could adopt a more sophisticated strategy along the lines of the Relaxed LASSO in [39]. Henceforth let $(\hat{\theta}_0(\lambda), \hat{\boldsymbol{\theta}}(\lambda))$ denote the regression coefficients obtained after the second step. Let $d_i(\lambda)$ denote the deviance of observation i under the model $(\hat{\theta}_0(\lambda), \hat{\boldsymbol{\theta}}(\lambda))$, i.e.

$$d_i(\lambda) = y_i(\hat{\theta}_0(\lambda) + \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\theta}}) - \log(1 + \exp(\hat{\theta}_0(\lambda) + \mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\theta}})).$$

Instead of the AIC and BIC we could use a robust measures of predictive error

$$-2 \operatorname*{median}_{i=1,\ldots,n} d_i(\lambda) + \frac{2}{n}\mathrm{df}(\hat{\boldsymbol{\theta}}(\lambda))$$

in lieu of AIC and

$$-2 \underset{i=1,\ldots,n}{\text{median}}\, d_i(\lambda) + \frac{\log(n)}{n}\text{df}(\hat{\boldsymbol{\theta}}(\lambda))$$

in lieu of BIC.

The reason we do this is because a robust fitting procedure will produce models under which outliers will have large deviances. We actually want to select models that correctly assign large deviances to outliers. Thus, the total deviance is an inappropriate measure of the prediction error if outliers were present. On the other hand, the median deviance, for example, would provide a more unbiased measure of the prediction error whether outliers are present or not. The final model selected would be the one that minimizes the robust prediction error criterion.

3.7 Proof of Theorem 2

It is immediate that $L_w(\boldsymbol{\theta};\boldsymbol{\theta}) = L_w(\boldsymbol{\theta})$. We turn our attention to proving that $L_w(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) \geq L_w(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \mathbb{R}^{p+1}$. Since $L_w(\boldsymbol{\theta})$ has bounded curvature our strategy is to represent $L_w(\boldsymbol{\theta})$ by its exact second order Taylor expansion about $\tilde{\boldsymbol{\theta}}$ and then find a tight uniform bound over the quadratic term in the expansion.

Recall by the first order Taylor expansion formula [37], if $f : U \subset \mathbb{R}^m \rightarrow \mathbb{R}$, twice differentiable at $\mathbf{x}_0$, then

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \sum_{i=1}^{m} h_i \frac{\partial f}{\partial x_i}(\mathbf{x}_0) + R_1(\mathbf{h}, \mathbf{x}_0),$$

where

$$R_1(\mathbf{h}, \mathbf{x}_0) = \sum_{i,j=1}^{m} \frac{1}{2} \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{c}_{ij}) h_i h_j,$$

and $\mathbf{c}_{ij} = \mathbf{x}_0 + \xi_{ij}\mathbf{h}$ for some $\xi_{ij} \in (0,1)$. In other words, $\mathbf{c}_{ij}$ is some point on the line segment connecting $\mathbf{x}_0$ and $\mathbf{x}_0 + \mathbf{h}$.

We now compute gradients and Hessians to construct an exact quadratic expansion of $L_w(\boldsymbol{\theta})$ around a point $\tilde{\boldsymbol{\theta}}$. Note that the loss has the following form.

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} f(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}),$$

where $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable. The chain rule gives the gradient

$$\nabla L(\boldsymbol{\theta}) = \sum_{i=1}^{n} f'(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta})\mathbf{x}_i = \mathbf{X}^\mathsf{T}\mathbf{v},$$

where $v_i = f'(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta})$. A second application of the chain rule gives the Hessian.

$$\nabla_2 L(\boldsymbol{\theta}) = \sum_{i=1}^{n} f''(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta})\mathbf{x}_i\mathbf{x}_i^\mathsf{T} = \mathbf{X}^\mathsf{T}\mathbf{M}\mathbf{X},$$

where M is a diagonal matrix with $m_{ii} = f''(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta})$. Set

$$f(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}) = \frac{w^2}{n}\left(F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta})^2 + (1 - F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}))^2\right) - 2\frac{w}{n}\left(y_i F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}) + (1 - y_i)(1 - F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\theta}))\right)$$

Then the gradient of $L_w(\boldsymbol{\theta})$ is $\nabla L_w(\boldsymbol{\theta}) = 2(w/n)\mathbf{X}^\mathsf{T}\mathbf{G}(w\mathbf{q} - \mathbf{u})$, and the Hessian of $L_w(\boldsymbol{\theta})$ is $\mathbf{H}_{\boldsymbol{\theta}} = 2(w/n)\mathbf{X}^\mathsf{T}\mathbf{M}_{\boldsymbol{\theta}}\mathbf{X}$, where $\mathbf{M}_{\boldsymbol{\theta}}$ is a diagonal matrix with diagonal entries $(\mathbf{M}_{\boldsymbol{\theta}})_{ii} = \psi_{u_i}(p_i)$ and where

$$\psi_u(p) = [2wp(1-p) - (2p-1)(w(2p-1) - u)]p(1-p).$$

Note that $\mathbf{q}$, $\mathbf{G}$, and $\mathbf{M}_{\boldsymbol{\theta}}$ depend on $\boldsymbol{\theta}$ through $\mathbf{p}$. Thus, $L_w(\boldsymbol{\theta})$ can be represented exactly as a second order Taylor expansion about $\tilde{\boldsymbol{\theta}}$:

$$L_w(\boldsymbol{\theta}) = L(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\mathsf{T}\nabla L_w(\tilde{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\mathsf{T}\mathbf{H}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where $\boldsymbol{\theta}^* = \tilde{\boldsymbol{\theta}} + \xi\boldsymbol{\theta}$ for some ξ between 0 and 1. Note that $(\mathbf{M}_{\boldsymbol{\theta}})_{ii}$ is bounded from above - i.e. $\sup_{\boldsymbol{\theta}\in\Theta}(\mathbf{M}_{\boldsymbol{\theta}})_{ii} < \infty$. We now introduce a surrogate function $L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$:

$$L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = L_w(\tilde{\boldsymbol{\theta}}) + 2\frac{w}{n}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{G}(w\mathbf{q} - \mathbf{u}) + \frac{w}{n}\eta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where $\mathbf{q}$, $\mathbf{u}$,and $\mathbf{G}$ are evaluated at $\tilde{\boldsymbol{\theta}}$ and

$$\eta \geq \max\left\{\sup_{p\in[0,1]} \psi_{-1}(p), \sup_{p\in[0,1]} \psi_1(p)\right\}.$$

We now argue that $L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$ majorizes $L_w(\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$. Note that for any $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$, $(\mathbf{M}_{\boldsymbol{\theta}})_{ii} \leq \eta$. Therefore,

$$(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{H}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{M}_{\boldsymbol{\theta}^*}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \leq \eta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

and consequently $L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$ majorizes $L_w(\tilde{\boldsymbol{\theta}})$ at $\tilde{\boldsymbol{\theta}}$.

The following observations lead to a simpler lower bound on η. Note, that.

$$\sup_{p\in[0,1]} \psi_{-1}(p) = \sup_{p\in[0,1]} \psi_1(p),$$

since $\psi_{-1}(p) = \psi_1(1-p)$. So, the lower bound on η can be more simply expressed as

$$\eta \geq \sup_{p\in[0,1]} \psi_1(p) = \max_{p\in[0,1]} \psi_1(p) = \frac{1}{4}\max_{a\in[-1,1]}\left\{\frac{3}{2}wa^4 - a^3 - 2wa^2 + a + \frac{w}{2}\right\}.$$

The first equality follows from the compactness of $[0,1]$ and the continuity of $\psi_1(p)$. The second equality follows from reparameterizing $\psi_1(p)$ in terms of $q = 2p - 1$. □

3.8 Derivation of Coordinate Descent Update Rules

We now provide details of the derivation of the coordinate descent update rules sketched in [21]. Recall that a subgradient of a convex function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ at a point $\mathbf{x}$ in the domain of f is a vector $\mathbf{g} \in \mathbb{R}^p$ such that $f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\mathsf{T}(\mathbf{y} - \mathbf{x})$ for any $\mathbf{y}$ in the domain of f. The subdifferential of a function f at $\mathbf{x}$ denoted $\partial f(\mathbf{x})$ is the set of subgradients of f at $\mathbf{x}$. If f is convex then $\mathbf{x}^*$ is a global minimizer of f if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$. This is immediate from the definitions since $f(\mathbf{y}) \geq f(\mathbf{x}^*) + \mathbf{0}^\mathsf{T}(\mathbf{y} - \mathbf{x}^*)$ for all $\mathbf{y}$ in the domain of f if and only if $\mathbf{0} \in \partial f(\mathbf{x}^*)$.

Consider minimizing the following strictly convex univariate function

$$f(x) = \frac{1}{2}(x - a)^2 + \lambda|x|, \tag{3.5}$$

where $\lambda \geq 0$. Note that since f is strictly convex it has at most one global minimizer. Since all the level sets of f are compact f has at least one global minimizer. Thus, there is a unique x^* such that $0 \in \partial f(x^*)$. The subdifferential is given by

$$\partial f(x) = \begin{cases} \{x - a + \lambda\} & \text{if } x > 0 \\ \{x - a - \lambda\} & \text{if } x < 0 \\ \{\lambda u - a : u \in [-1, 1]\} & \text{if } x = 0 \end{cases}$$

Let us go through each possible case for x^*. Note that $x^* > 0 \Leftrightarrow x^* - a + \lambda = 0 \Leftrightarrow x^* = a - \lambda > 0 \Leftrightarrow a > \lambda$. Thus, if $a > \lambda$ then the global minimizer x^* is $a - \lambda$. By a similar argument, if $a < -\lambda$ then the global minimizer x^* is $a + \lambda$. Finally, consider the case that $x^* = 0$. Note that $x^* = 0 \Leftrightarrow 0 \in \partial f(0) \Leftrightarrow$ there is a $u^* \in [-1, 1]$ such that $\lambda u^* - a = 0$. Therefore, if $-\lambda \leq a \leq \lambda$ the global minimizer x^* is 0. Putting all three cases together gives $x^* = S(a, \lambda)$ is the unique global minimizer of (3.5).

We are now ready to derive the coordinate descent update for the Elastic Net penalized least squares problem. Let $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\theta} \in \mathbb{R}^p$. Define the partial response

$$\mathbf{y}^{(-k)} = [\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{0}, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_p]\boldsymbol{\theta},$$

and consider the kth coordinate update problem

$$\begin{aligned}
\hat{\theta}_k =& \arg\min_{\theta_k} \left\{ \frac{1}{2} L(\mathbf{y}, \mathbf{X}\boldsymbol{\theta}) + \lambda J(\boldsymbol{\theta}) \right\} \\
=& \arg\min_{\theta_k} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \left(\alpha \|\boldsymbol{\theta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\theta}\|_2^2 \right) \right\} \\
=& \arg\min_{\theta_k} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - y_i^{(-k)} - x_{ik}\theta_k)^2 + \lambda \frac{1-\alpha}{2} \theta_k^2 + \lambda\alpha|\theta_k| \right\} \\
=& \arg\min_{\theta_k} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left[x_{ik}^2 \theta_k^2 - 2 \left((y_i - y_i^{(-k)}) x_{ik} \right) \theta_k \right] + \lambda \frac{1-\alpha}{2} \theta_k^2 + \lambda\alpha|\theta_k| \right\} \\
=& \arg\min_{\theta_k} \left\{ \frac{1}{2} \left(\left[\sum_{i=1}^{n} x_{ik}^2 + \lambda(1-\alpha) \right] \theta_k^2 - 2 \left(\sum_{i=1}^{n} x_{ik}(y_i - y_i^{(-k)}) \right) \theta_k \right) + \lambda\alpha|\theta_k| \right\} \\
=& \arg\min_{\theta_k} \left\{ \frac{1}{2} \left(\theta_k - \frac{\sum_{i=1}^{n} x_{ik}(y_i - y_i^{(-k)})}{\sum_{i=1}^{n} x_{ik}^2 + \lambda(1-\alpha)} \right)^2 + \frac{\lambda\alpha}{\sum_{i=1}^{n} x_{ik}^2 + \lambda(1-\alpha)} |\theta_k| \right\} \\
=& S\left(\frac{\sum_{i=1}^{n} x_{ik}(y_i - y_i^{(-k)})}{\sum_{i=1}^{n} x_{ik}^2 + \lambda(1-\alpha)}, \frac{\lambda\alpha}{\sum_{i=1}^{n} x_{ik}^2 + \lambda(1-\alpha)} \right) \\
=& \frac{S(\sum_{i=1}^{n} x_{ik}(y_i - y_i^{(-k)}), \lambda\alpha)}{\sum_{i=1}^{n} x_{ik}^2 + \lambda(1-\alpha)}.
\end{aligned}$$

CHAPTER 4

GLOBAL CONVERGENCE OF THE LOGISTIC L_2E ALGORITHM

In this Chapter we prove that the MM algorithm described in Chapter 3 generates iterates that converge to a stationary point. We will require MM algorithm convergence results for locally Lipschitz objective functions. While more general results can be found in the literature [44], in this Chapter we prove a version of global convergence we need for our algorithm.

After reviewing some concepts upon which the proofs are based we will proceed as follows. We first prove a general theorem for convergence of iterative algorithms for the minimization of non-smooth functions so that we can apply it to the algorithms described in Chapter 3. Specifically, we will prove convergence to a stationary point for any iterative algorithm to minimize a continuous but possibly non-differentiable objective function under suitable regularity conditions. We will then use this general result to prove the convergence of a general MM algorithm when the objective function is locally Lipschitz continuous. Finally, we prove the convergence of the coordinate descent version of the MM algorithm when used to minimize the Elastic Net penalized L_2E logistic loss.

4.1 Analysis for Optimization

A key condition in the convergence proofs to follow is coerciveness since it is a sufficient condition to ensure the existence of a global minimum.

Definition 3. *A continuous function* $f : U \subset \mathbb{R}^n \to \mathbb{R}$ *is coercive if all its level sets* $S_t = \{\mathbf{x} \in U : f(\mathbf{x}) \leq t\}$ *are compact.*

It is not hard to show that for a coercive function f, that $f \to \infty$ whenever $\|x\|_2 \to \infty$. Intuitively we expect that such a function which "blows" up at its "boundaries" will attain its minimum. To show this formally take any $\mathbf{z} \in U$. The set $S_{f(\mathbf{z})} = \{\mathbf{x} \in U : f(\mathbf{x}) \leq f(\mathbf{z})\}$ is compact. Since f is continuous it attains its lower bound $f(\mathbf{x}^*)$ on $S_{f(\mathbf{z})}$. Note that $S_{f(\mathbf{z})} \subset S_{f(\mathbf{z}')}$ for any $\mathbf{z}' \in U$ such that $f(\mathbf{z}') \geq f(\mathbf{z})$. Therefore $f(\mathbf{x}^*)$ is the global minimum.

We reiterate that coerciveness is a sufficient but not necessary condition for functions to attain a minimum. The function $1 - \exp(-|x|)$ has a global minimum at 0 but is not coercive.

Note that a coercive function does not need to be differentiable. Indeed, we are interested in minimizing nonsmooth objective functions, i.e. functions which may not be differentiable. Nonetheless, our ambitions are still relatively modest; the penalized loss functions we wish to minimize are differentiable everywhere except at a single point. In other words, the objective functions of interest are differentiable almost everywhere except on a set of Lebesgue measure zero.

The set of functions that are almost everywhere differentiable over open sets of Euclidean space include locally Lipschitz continuous functions (Rademacher's Theorem). This is important to note since much of optimization theory that was developed under smoothness assumptions can be generalized to nonsmooth settings, specifically when objective functions are locally Lipschitz continuous. The following definitions and properties

are discussed in greater detail in Clarke [12]. We review the definitions of Lipschitz continuity and local Lipschitz continuity.

Definition 4. *Let U_0 be an open subset of $\mathbb{R}^n$. A function $f : U \to \mathbb{R}$ is* Lipschitz *on U_0 if there exists a $K > 0$ such that $|f(\mathbf{x}) - f(\mathbf{y})| \leq K\|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in U_0$.*

Definition 5. *A function f is* locally Lipschitz *if for all $\mathbf{x}$ there exists a neighborhood $N(\mathbf{x})$ of $\mathbf{x}$ on which it is Lipschitz. In other words there exists a $K_{\mathbf{x}} > 0$ such that $|f(\mathbf{u}) - f(\mathbf{v})| \leq K_{\mathbf{x}}\|\mathbf{u} - \mathbf{v}\|_2$ for all $\mathbf{u}, \mathbf{v} \in N(\mathbf{x})$.*

Convex functions are Lipschitz except for pathological cases. See Proposition 2.2.6 and its Corollary in [12].

We will use the following three facts to prove our algorithm converges. The proofs are straightforward and undoubtedly not new, but for completeness proofs are given in Section 4.4.

Proposition 4.1.0.1. *Finite sums of locally Lipschitz continuous functions are also locally Lipschitz continuous.*

Proposition 4.1.0.2. *Convex quadratic functions are locally Lipschitz continuous.*

Proposition 4.1.0.3. *The mapping $\mathbf{x} \to \|\mathbf{x}\|_1$ is a locally Lipschitz continuous function.*

Locally Lipschitz functions are not necessarily differentiable, but there are weaker and more general notions of differentiability and gradients that can be made for locally Lipschitz functions. We need these generalized notions of gradients to define stationary points of locally Lipschitz functions.

Definition 6. *Let U_0 be an open subset of $\mathbb{R}^n$. If $f : U_0 \to \mathbb{R}$ is locally Lipschitz at $\mathbf{x} \in U_0$, the* generalized directional derivative *of f at $\mathbf{x} \in U_0$ in the direction of $\mathbf{v} \in \mathbb{R}^n$, denoted $f^\circ(\mathbf{x}; \mathbf{v})$, is given by*

$$f^{\circ}(\mathbf{x};\mathbf{v}) = \limsup_{\mathbf{y}\to\mathbf{x},t\downarrow 0} \frac{f(\mathbf{y}+t\mathbf{v}) - f(\mathbf{y})}{t}.$$

Definition 7. *The* generalized gradient *of f at $\mathbf{x} \in U_0$, denoted $\partial f(\mathbf{x})$, is defined as follows:*

$$\partial f(\mathbf{x}) = \{\boldsymbol{\xi} \in \mathbb{R}^n : f^{\circ}(\mathbf{x};\mathbf{v}) \geq \boldsymbol{\xi}^{\mathsf{T}}\mathbf{v} \text{ for all } \mathbf{v} \in \mathbb{R}^n\}.$$

The generalized gradient is always non-empty for locally Lipschitz functions. In fact it is a nonempty, convex, compact subset of $\mathbb{R}^n$. See Proposition 2.1.2 in [12]. Note that we have overloaded the notation for subdifferentials in Chapter 3 and generalized gradients. This was intentional since the generalized gradient of a convex function is its subdifferential. See Proposition 2.2.7 in [12]. Moreover, when f is differentiable, $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$,

Finally, we will use the following definition of stationary points for locally Lipschitz functions in the subsequent discussion. Note that when f is smooth or convex, we recover the appropriate definitions of stationary points.

Definition 8. *A point $\mathbf{x} \in U_0$ such that $\mathbf{0} \in \partial f(\mathbf{x})$ is called a* stationary point *of f.*

Figure 4.1 shows a locally Lipschitz continuous function; it is in fact a Lipschitz function on $[-2, 7]$. The generalized gradient for several points marked by A, B, C, and D are given by $\partial f(-1/2) = \{-\pi/\sqrt{2}\}, \partial f(2) = [0, 2], \partial f(2.5) = [-1, 2], \partial f(5) = [-1, 1]$. The set of stationary points is given by $\{2\} \cup [0, 2] \cup \{2.5\} \cup \{5\}$.

4.2 Convergence of General Iterative Minimization Algorithms

Let f map an open convex subset U of $\mathbb{R}^n$ to the reals. Let M be an iteration function that maps points in U into U. The function M generates the best minimizer of f based on

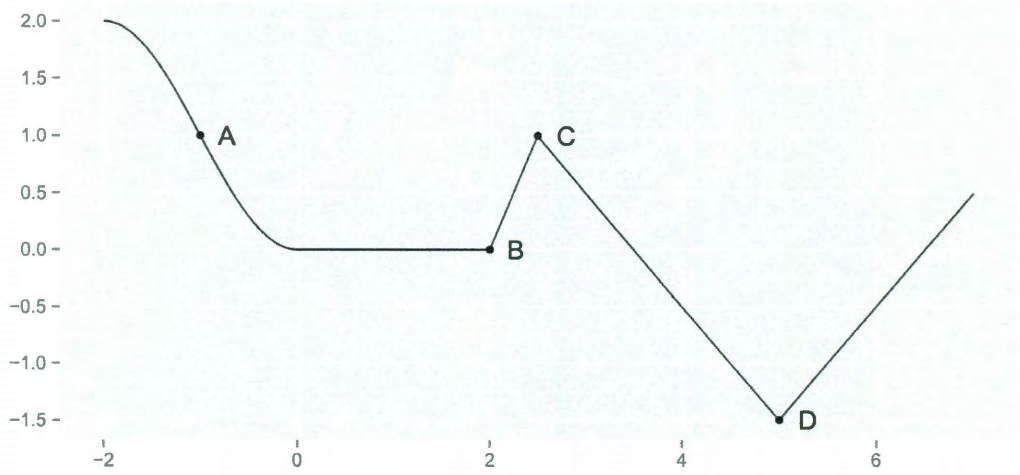


Figure 4.1: A locally Lipschitz continuous function with several stationary points.

its previous best minimizer. In this section we discuss sufficient conditions for the iterates generated by M to converge to a stationary point of f.

Suppose we have the following conditions on an objective function f.

B1: f is coercive.

B2: f has finitely many stationary points.

Suppose the iteration map $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ has the following properties:

B3: M is continuous

B4: $\mathbf{y}$ is a fixed point of M if and only if it is a stationary point of f.

B5: $f(M(\mathbf{y})) \leq f(\mathbf{y})$, with equality if and only if $M(\mathbf{y}) = \mathbf{y}$.

These conditions are almost exactly the same as those listed in [33] for global convergence of block relaxation and the MM algorithm when the objective and majorizing functions are smooth. The difference is that we require that the objective function has finitely many stationary points, whereas [33] requires that the objective function has only isolated stationary points. The purpose of requiring isolated stationary points in [33], however, is to establish that the objective function has only finitely many stationary points under differentiability assumptions. Because we do not make any assumptions about the differentiability

of functions we have to explicitly require that the objective function has only finitely many stationary points.

We will need two adaptations of propositions in [33] (Propositions 15.4.1 and 15.5.2) that characterize the limit points of the sequence of iterates when B1-B5 hold. For completeness proofs are given here to make explicit the roles the various assumptions play in the proof; the arguments are essentially the same as in [33]

Proposition 4.2.0.4. *Suppose B1, B3-B5 are true. Then the sequence* $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ *for any initial* $\mathbf{x}_1$ *has at least one limit point and all its limit points are stationary points of* f.

Proof. Since f is coercive (B1), the set $S = \{\mathbf{x} \in U : f(\mathbf{x}) \leq f(\mathbf{x}_1)\}$ is compact. Since the sequence $\mathbf{x}_n$ is contained in S, it has a convergent subsequence, $\mathbf{x}_{n_k}$. Let $\mathbf{z}$ denote its limit, i.e., $\mathbf{z} = \lim_{k\to\infty} \mathbf{x}_{n_k}$. Our goal is to show that $f(\mathbf{z}) = f(M(\mathbf{z}))$ since $\mathbf{z}$ is a fixed point of M if and only if it is a stationary point according to B4.

The sequence $f(\mathbf{x}_n)$ is non-increasing and is bounded below since f has a global minimum which follows from f being coercive and continuous. Therefore, $\lim_{n\to\infty} f(\mathbf{x}_n)$ exists.

If a sequence converges to a limit, then all its subsequences must converge to the same limit. Therefore, $\lim_{k\to\infty} f(\mathbf{x}_{n_k}) = \lim_{n\to\infty} f(\mathbf{x}_n)$. Since f is continuous we can simplify the left limit to get $f(\mathbf{z}) = \lim_{n\to\infty} f(\mathbf{x}_n)$. If we can show that $\lim_{n\to\infty} f(\mathbf{x}_n) = f(M(\mathbf{z}))$ then we are done. But this follows since

$$\lim_{n\to\infty} f(\mathbf{x}_n) = \lim_{n\to\infty} f(\mathbf{x}_{n+1}) = \lim_{n\to\infty} f(M(\mathbf{x}_n)) = \lim_{k\to\infty} f(M(\mathbf{x}_{n_k})) = f(M(\mathbf{z})).$$

The third equality follows from the fact that $f(M(\mathbf{x}_n))$ is a bounded non-increasing sequence and thus a convergent sequence. The fourth equality follows from the fact that subsequences of convergent sequences must converge to the same limit as the original convergence sequence. The last equality follows from the continuity of f and M (B3). There-

fore, $f(\mathbf{z}) = \lim_{n\to\infty} f(\mathbf{x}_n) = f(M(\mathbf{z}))$. By B5, $\mathbf{z}$ is a fixed point of M, and, therefore, by B4, $\mathbf{z}$ is a stationary point of f. □

The previous proposition proves that any sequence of iterates has at least one limit point and that every limit point is a stationary point. To make the stronger claim that the iterates converge to a stationary point, we need to show that every iterate sequence is bounded and every iterate sequence has exactly one limit point. The latter is proven by showing that the set of limit points for an iterate sequence is connected. We will need the following proposition which will be used in turn to prove that the set of limit points of the iteration sequence $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ is a connected.

Proposition 4.2.0.5. *If a bounded sequence* $\mathbf{x}_m$ *in* $\mathbb{R}^n$ *satisfies* $\lim_{m\to\infty} \|\mathbf{x}_{m+1} - \mathbf{x}_m\|_2 = 0$, *then its set* Γ *of limit points is connected.*

Proposition 4.2.0.5 is Proposition 8.2.1 in [33] and a proof can be found in that reference. The previous proposition is used to show that the limit points of the iterates generated by M from any starting point are connected.

Proposition 4.2.0.6. *Suppose B1, B3-B5 are true. Then the set* Γ *of limit points of the sequence* $\mathbf{b}_{n+1} = M(\mathbf{b}_n)$ *is connected.*

Proposition 4.2.0.6 is Proposition 15.4.2 in [33] and a proof can be found in that reference. We are now ready to prove the convergence result for general iterative minimization algorithms.

Lemma 3. *Suppose* $B1$-$B5$ *are true, then the sequence* $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ *starting from any point* $\mathbf{x}_1$ *will converge to a stationary point of* f.

Proof. A finite non-empty connected set contains exactly one point. Thus, if f has finitely many stationary points, then the sequence $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ has exactly one limit point which is a fixed point and therefore by B4 a stationary point. Note that the sequence

$\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ is bounded. A bounded sequence with exactly one limit point converges to that point. Therefore, regardless of starting position the sequence of iterates generated by M converges to a stationary point of f. □

If B1, B3-B5 are true but B2 is not then all we can say is that the limit points of any iterate sequence are stationary points. The set of limit points are still connected however. So, even if the sequence of iterates does not converge, it will get arbitrarily close to its connected set of limit points.

Proposition 4.2.0.7. *Suppose B1, B3-B5 are true, then the sequence* $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ *starting from any point* $\mathbf{x}_1$ *will get arbitrarily close to a connected subset of stationary points of* f.

Proof. Let Γ denote the set of limit points of a sequence $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ starting from some point $\mathbf{x}_1$. By Proposition 4.2.0.6 and 4.2.0.4, the set Γ is a connected subset of stationary points of f. Fix $\epsilon > 0$. Let $T_\epsilon = \bigcup_{\mathbf{g}\in\Gamma} B(\mathbf{g}, \epsilon)$. We will show that only finitely many $\mathbf{x}_n \notin T_\epsilon$.

Note that $\Omega = T_\epsilon^c \cap S_{f(\mathbf{x}_1)}$ is a closed subset of the compact set $S_{f(\mathbf{x}_1)}$ and is therefore compact. Suppose $\mathbf{x}_n \in \Omega$ infinitely often. Since Ω is compact, then $\mathbf{x}_n$ has a limit point in Ω. But Γ contains all the limit points of $\mathbf{x}_n$, and $\Gamma \cap \Omega = \emptyset$. Therefore, only finitely many $\mathbf{x}_n \in \Omega$. □

4.2.1 Convergence MM Algorithms for Locally Lipschitz Functions

We will use Lemma 3 to prove that an MM algorithm converges to a stationary point. Let $U \subset \mathbb{R}^n$, our parameter space, be a convex set. Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ denote the objective function to be minimized and $g : U \times U \to \mathbb{R}$ a majorization of f. Let $M(\mathbf{x}) \in \arg\min_{\mathbf{y}\in U} g(\mathbf{y}|\mathbf{x})$ be a point-to-set mapping defining the MM updates. We will prove that any sequence of iterates generated by M will converge to a stationary point of f under the following regularity conditions.

A1. The objective function f is coercive.
A2. f is locally Lipschitz continuous.
A3. The set of stationary points of f is finite.
A4. The majorizer g is strict.
A5. $g(\mathbf{b}|\mathbf{a})$ is strictly convex in $\mathbf{b}$.
A6. $g(\mathbf{b}|\mathbf{a})$ is continuous on $U \times U$ and locally Lipschitz in $\mathbf{b}$.

Note that none of the conditions require either f or g to be differentiable. This is important since the Elastic Net penalty function is *not* differentiable due to the LASSO term. The proof that A1-A6 guarantee convergence of the iterates to a stationary point relies on a result in Schifano el al. [44]. Specifically, we use results from Schifano el al. to show that A4-A6 imply B3 and B4. We use Proposition A.8 in [44] which is restated here using notation used in this dissertation.

Proposition 4.2.1.1 ((Proposition A.8 in [44])). *Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ denote the objective function to be minimized and $g : U \times U \to \mathbb{R}$ a majorization of f. Let $M(\mathbf{x}) \in \arg\min_{\mathbf{y} \in U} g(\mathbf{y}|\mathbf{x})$ be a point-to-set mapping defining the MM updates. Suppose the following regularity conditions are true.*

R1. The objective function $f(\mathbf{b})$ is locally Lipschitz continuous for $\mathbf{b} \in U$ and there exists at least one $\mathbf{b}_0 \in U$ such that $L(f(\mathbf{b}_0)) = \{\mathbf{b} \in U : f(\mathbf{b}) \leq f(\mathbf{b}_0)\}$ is compact. Assume that the set of stationary points $\mathcal{S}$ of $f(\mathbf{b})$ is a finite set.
R2. $f(\mathbf{b}) = g(\mathbf{b}|\mathbf{b})$ for each $\mathbf{b} \in U$.
R3. The majorization g is strict: $g(\mathbf{b}|\mathbf{a}) > g(\mathbf{b}|\mathbf{b})$ for $\mathbf{b} \neq \mathbf{a}$, for all $\mathbf{b}, \mathbf{a} \in U$.
R4. $g(\mathbf{b}|\mathbf{a})$ is continuous for $(\mathbf{a}, \mathbf{b}) \in U \times U$ and locally Lipschitz in $\mathbf{b}$ for $\mathbf{b}$ near $\mathbf{a}$.
R5. $M(\mathbf{b})$ is a singleton set consisting of one bounded vector for each $\mathbf{b} \in U$.

Then, a point is a fixed point of M if and only if it is a stationary point of f.

We now show that an MM algorithm that satisfies A1-A6 also satisfies R1-R5.

Proposition 4.2.1.2. *The conditions A1-A6 imply R1-R5.*

Proof. Take any $\mathbf{z} \in U$. Because f is coercive, $L(f(\mathbf{z}))$ is compact. Thus, A1-A3 imply R1. Condition R2 and R3 follow immediately from A4. Condition A6 implies R4. Note that $B = \{\mathbf{b} \in U : g(\mathbf{b}|\mathbf{a}) \leq g(\mathbf{a}|\mathbf{a}) = f(\mathbf{a})\}$ is a closed subset of $C = \{\mathbf{b} \in U : f(\mathbf{b}) \leq f(\mathbf{a})\}$. B is closed because $g(\mathbf{b}|\mathbf{a})$ is locally Lipschitz continuous in $\mathbf{b}$ and therefore continuous in $\mathbf{b}$. Note that $f(\mathbf{b}) \leq g(\mathbf{b}|\mathbf{a})$ since g majorizes f. So, if $g(\mathbf{b}|\mathbf{a}) \leq f(\mathbf{a})$ then $f(\mathbf{b}) \leq g(\mathbf{b}|\mathbf{a}) \leq f(\mathbf{a})$. Therefore, B is a subset of C and is consequently compact. Since $g(\mathbf{b}|\mathbf{a})$ is continuous in $\mathbf{b}$, $g(\cdot|\mathbf{a})$ achieves its global minimum on B. Moreover, since $\min_{\mathbf{b}\in U} g(\mathbf{b}|\mathbf{a}) = \min\{\min_{\mathbf{b}\in B} g(\mathbf{b}|\mathbf{a}), \min_{\mathbf{b}\in B^c} g(\mathbf{b}|\mathbf{a})\} = \min_{\mathbf{b}\in B} g(\mathbf{b}|\mathbf{a})$, the function $g(\cdot|\mathbf{a})$ achieves its global minimum on U. Since $g(\mathbf{b}|\mathbf{a})$ is strictly convex in $\mathbf{b}$ it has a unique global minimizer in U. Therefore, condition R5 is met as well. □

We are now ready to show prove the convergence of MM algorithms to stationary points when the objective and majorization are locally Lipschitz continuous.

Theorem 4. *Let $U \subset \mathbb{R}^n$ be a convex set. Let $f : U \subset \mathbb{R}^n \to \mathbb{R}$ denote the objective function to be minimized and $g : U \times U \to \mathbb{R}$ be a majorization of f. Let $M(\mathbf{x}) \in \arg\min_{\mathbf{y}\in U} g(\mathbf{y}|\mathbf{x})$. Suppose conditions A1-A6 are met. Then the iterate sequence $\mathbf{x}_{n+1} = M(\mathbf{x}_n)$ converges to a stationary point of f.*

Proof. Note A1-A6 meet conditions B1-B5. A1 is B1 and A3 is B2. Note that R4 and R5 imply that M is continuous. R4 follows from A6 and R5 follows from A1, A5 and A6. Thus, A1, A5, and A6 imply B3. B4 follows from Proposition A.8 in Schifano. B5 follows from A4, the requirement that the majorization be strict. If $M(\mathbf{y}) = \mathbf{y}$ then $f(M(\mathbf{y})) = f(\mathbf{y})$. If $M(\mathbf{y}) \neq \mathbf{y}$, then

$$f(M(\mathbf{y})) \leq g(M(\mathbf{y})|\mathbf{y}) < g(\mathbf{y}|\mathbf{y}) = f(\mathbf{y}).$$

□

4.3 Convergence of the MM algorithm of the Elastic Net Penalized L_2E Logistic Loss

Consider the regularized loss:

$$L_{w,\lambda,\alpha}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{w^2 \sum_{y\in\{0,1\}} P_{\boldsymbol{\theta}}(Y_i = y|\mathbf{x}_i)^2 - 2wP_{\boldsymbol{\theta}}(Y_i = y_i|\mathbf{x}_i)\right\} + \lambda\left(\alpha\|\boldsymbol{\theta}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{\theta}\|_2^2\right),$$
$$= \frac{1}{n}\|\mathbf{t}_{\boldsymbol{\theta}}(\mathbf{X}) - \mathbf{y}\|_2^2 - w\overline{y} + \lambda\left(\alpha\|\boldsymbol{\theta}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{\theta}\|_2^2\right),$$

where

$$\mathbf{t}_{\boldsymbol{\theta}}(\mathbf{X}) = w\left(P_{\boldsymbol{\theta}}(Y_1 = 1|\mathbf{x}_1), \ldots, P_{\boldsymbol{\theta}}(Y_n = 1|\mathbf{x}_n), P_{\boldsymbol{\theta}}(Y_1 = 0|\mathbf{x}_1), \ldots, P_{\boldsymbol{\theta}}(Y_n = 0|\mathbf{x}_n)\right)^{\mathsf{T}}$$

and

$$\mathbf{y} = (y_1, \ldots, y_n, 1-y_1, \ldots, 1-y_n)^{\mathsf{T}}$$

Since $\mathbf{t}_{\boldsymbol{\theta}}(\mathbf{X})$ is bounded below by $-w\overline{y}$ and $(1/2)\|\boldsymbol{\theta}\|_2^2$ is coercive, $L_{w,\lambda,\alpha}(\boldsymbol{\theta})$ is coercive. A1 is met.

Note that the gradient of the $L_w(\boldsymbol{\theta})$ is $\nabla L_w(\boldsymbol{\theta}) = 2(w/n)\mathbf{X}^{\mathsf{T}}\mathbf{G}(w\mathbf{q} - \mathbf{u})$. The norm of the gradient is bounded; specifically it is no greater than $w(w+1)\sigma_1^2/2$ where σ_1 is the largest singular value of $\mathbf{X}$. Therefore, $L_w(\boldsymbol{\theta})$ is Lipschitz continuous and therefore locally Lipschitz continuous. Consequently, $L_{w,\lambda,\alpha}(\boldsymbol{\theta})$ is locally Lipschitz continuous. A2 is met.

Recall the majorization we are using is given by

$$L_w(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = L_w(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\nabla L_w(\tilde{\boldsymbol{\theta}}) + \frac{w}{n}\eta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}),$$

where

$$\eta > \frac{1}{4}\max_{a\in[-1,1]}\left\{\frac{3}{2}wa^4 - a^3 - 2wa^2 + a + \frac{w}{2}\right\}.$$

Note to ensure that the majorization is strict we need the inequality to be strict. Thus, the curvature of the majorization exceeds the maximum curvature of L_w and the majorization is strict. A4 is met.

Note that $L_w(\boldsymbol{\theta}) + (1/2)\lambda(1-\alpha)\|\boldsymbol{\theta}\|_2^2$ is strictly convex if $\lambda(1-\alpha) > 0$. The sum of a strictly convex function with a convex function is strictly convex. So, A5 is met.

The penalized majorization is the sum of continuous functions in $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \in U \times U$ and is consequently continuous. The penalized majorization as a function of its first argument is the sum of a positive definite quadratic function and the 1-norm function, both of which are locally Lipschitz continuous so their sum is locally Lipschitz continuous. A6 is met.

Thus, Algorithm 3 will converge to a stationary point of $L_{w,\lambda,\alpha}(\boldsymbol{\theta})$, provided that there are only finitely many stationary points and the coordinate descent minimization of the Elastic Net penalized quadratic majorization is solved exactly. The latter condition is met as discussed next.

Note that the iterate update $\mathbf{x}_{n+1} \leftarrow M(\mathbf{x}_n)$ can be accomplished by any means algorithmically so long as the global minimum of the majorization is found. We next show that applying coordinate descent on the penalized majorization will find the global minimum. We use Lemma 3.1 and Theorem 5.1 of [54]. Let f_0 be $L_w(\boldsymbol{\theta}) + (1/2)\lambda(1-\alpha)\|\theta\|_2^2$ and $f_i = \lambda\alpha|\theta_i|$. It can be verified that $f, f_0, f_1, \ldots, f_p$ are continuous and convex, and f is hemivariate. Thus assumptions B1-B3 are met. Moreover, f_0 satisfies Assumptions A1 in [54], since the domain is $\mathbb{R}^p$ and it is differentiable everywhere on its domain. Consequently, the sequence of iterates generated by the coordinate descent algorithm are bounded and the sequence's set of limit points are stationary points of f. Recall that f, however, is strictly convex and coercive and consequently has exactly one stationary point which is the global minimum of f.

4.4 Proofs

Proposition 4.4.0.3. *Finite sums of locally Lipschitz continuous functions are also locally Lipschitz continuous.*

Proof. We prove that the sum of two locally Lipschitz continuous functions is locally Lipschitz continuous. An induction argument would complete the proof. Suppose f and g are locally Lipschitz continuous. Fix $\mathbf{x}$. There are constants $K_{\mathbf{x}}$ and $K'_{\mathbf{x}}$ and neighborhoods $N_f(\mathbf{x})$ and $N_g(\mathbf{x})$ such that

$$\|f(\mathbf{u}) - f(\mathbf{v})\|_2 \leq K_{\mathbf{x}}\|\mathbf{u} - \mathbf{v}\|_2,$$

for all $\mathbf{u}, \mathbf{v} \in N_f(\mathbf{x})$. Similarly,

$$\|g(\mathbf{u}) - g(\mathbf{v})\|_2 \leq K'_{\mathbf{x}}\|\mathbf{u} - \mathbf{v}\|_2,$$

for all $\mathbf{u}, \mathbf{v} \in N_g(\mathbf{x})$. Let $N(\mathbf{x}) = N_f(\mathbf{x}) \cap N_g(\mathbf{x})$. Then for any $\mathbf{u}, \mathbf{v} \in N(\mathbf{x})$,

$$\begin{aligned}\|f(\mathbf{u}) + g(\mathbf{u}) - f(\mathbf{v}) - g(\mathbf{v})\|_2 &\leq \|f(\mathbf{u}) - f(\mathbf{v})\|_2 + \|g(\mathbf{u}) - g(\mathbf{v})\|_2 \\ &\leq (K'_{\mathbf{x}} + K_{\mathbf{x}})\|\mathbf{u} - \mathbf{v}\|_2.\end{aligned}$$

□

Proposition 4.4.0.4. *Convex quadratic functions are locally Lipschitz continuous.*

Proof. Let $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is positive semidefinite. Fix $\mathbf{x}$. Let $N(\mathbf{x})$ be the open ball of unit norm radius centered on $\mathbf{x}$. Take any $\mathbf{u}, \mathbf{v} \in N(\mathbf{x})$. Let $g : [0, 1] \to \mathbb{R}$ be defined as $g(t) = f(t\mathbf{u} + (1 - t)\mathbf{v})$. Then there is a $c \in (0. 1)$ such that

$$\begin{aligned}f(\mathbf{u}) - f(\mathbf{v}) = g(1) - g(0) = g'(c) &= \nabla f(c\mathbf{u} + (1 - c)\mathbf{v})^\mathsf{T}(\mathbf{u} - \mathbf{v}) \\ &\leq \|\nabla f(c\mathbf{u} + (1 - c)\mathbf{v})\|_2\|\mathbf{u} - \mathbf{v}\|_2.\end{aligned}$$

Hence,

$$\|f(\mathbf{u}) - f(\mathbf{v})\|_2 \leq \|\nabla f(c\mathbf{u} + (1-c)\mathbf{v})\|_2 \|\mathbf{u} - \mathbf{v}\|_2.$$

Note that

$$\begin{aligned}
\|\nabla f(c\mathbf{u}) + (1-c)\mathbf{v})\|_2 &\leq \sup_{\|\mathbf{w}-\mathbf{x}\|_2 \leq 1} 2\|\mathbf{A}\mathbf{w}\|_2 \\
&\leq \sup_{\|\mathbf{z}\|_2 \leq 1} 2\|\mathbf{A}\mathbf{z}\|_2 + 2\|\mathbf{A}\mathbf{x}\|_2 \\
&= 2\sigma_1 + 2\|\mathbf{A}\mathbf{x}\|_2,
\end{aligned}$$

where σ_1 is the largest eigenvalue of $\mathbf{A}$. Therefore for all $\mathbf{u}, \mathbf{v} \in N(\mathbf{x})$,

$$\|f(\mathbf{u}) - f(\mathbf{v})\|_2 \leq (2\sigma_1 + 2\|\mathbf{A}\mathbf{x}\|_2)\|\mathbf{u} - \mathbf{v}\|_2.$$

□

Proposition 4.4.0.5. *The mapping $\mathbf{x} \to \|\mathbf{x}\|_1$ is a locally Lipschitz continuous function.*

Proof. Let $f(\mathbf{x}) = \sum_i |x_i|$.

$$\left| \sum_i |x_i| - \sum_i |y_i| \right| \leq \sum_i \big| |x_i| - |y_i| \big| \leq \|\mathbf{x} - \mathbf{y}\|_1 \leq \sqrt{n}\|\mathbf{x} - \mathbf{y}\|_2$$

□

CHAPTER 5

SIMULATIONS

In this section we report on three simulations comparing the MLE and L_2E parameter estimates. The first two simulations examine the accuracy of estimation. We then follow with a simulation experiment designed to examine the variable selection properties. For the first two simulations we generated 1000 data sets, with 200 binary outcomes each associated with 4 covariates, from the logistic model specified in equation (2.3) with $\boldsymbol{\theta} = (0, 1, 0.5, 1, 2)^{\mathsf{T}}$. The covariates $\mathbf{x}_i$ were drawn from one of two populations. For $i = 1, \ldots, 100$, $x_{ij} = 0.25 + 0.4\epsilon_{ij}$ for $j = 1, 2, 3, 4$, and for $i = 101, \ldots, 200$, $x_{ij} = -0.25 + 0.4\epsilon_{ij}$, where ϵ_{ij} were iid $N(0, 1)$. Again for all i, $x_{i0} = 1$.

5.1 Varying the Location of a Single Outlier

In the first scenario, we added a single outlier, $(y_{201}, \mathbf{x}_{201})$ where $y_{201} = 0$ and $x_{201} = (1, \delta, \delta, \delta, \delta)^{\mathsf{T}}$ and δ took on values in $\{-0.25, 1.5, 3, 6, 12, 24\}$. We then performed logistic regression and L_2E regression with $w = 1$. Table 5.1 shows the mean and standard deviation for the fitted coefficient values.

The 201st point is being moved in covariate space along the line that runs through the centroids of the two subpopulations. There are three things to note. The MLE becomes

increasingly biased towards zero as the 201st point is moved from -0.25 to to 24. In contrast, the L_2E is insensitive to the placement of the 201st point. Figure 5.1 shows how $||\boldsymbol{\theta}||_2$ under each estimation procedure varies as the position of outlier is moved[1]. We see that MLE values demonstrate "implosion" breakdown, i.e. $||\boldsymbol{\theta}||_2$ tends towards 0 as the leverage of the 201st point increases. The L_2E estimates do not. The second observation is the cost of the L_2E's unbiasedness is increased variance as seen in the increased standard error in Table 5.1. The L_2E's sample standard error is greater than the MLE's for all locations of the outlier. The third observation is that the L_2E regression coefficients actually appear to be slightly biased away from zero when $w = w_{opt} = 200/201$.

5.2 Varying the Number of Outliers at a Fixed Location

In the second scenario, we add a variable number of outliers at a single location: $\{(y_i, \mathbf{x}_i)\}_{i=201}^{N}$, where $y_i = 0$ and $\mathbf{x}_i = (1, 3, 3, 3, 3)^{\mathsf{T}}$ for $i = 201, \ldots, N$ and the number of outliers $N = 0, 1, 5, 10, 15, 20$. Again we generated $1,000$ such data sets. Table 5.2 shows the mean and standard deviation for the fitted coefficient values.

We make the same three observations as before. Regardless of the number of outliers, the L_2E remains unbiased whereas the MLE does not (Table 5.2). Figure 5.2 shows how $||\boldsymbol{\theta}||_2$ under each estimation procedure varies as the number of outliers. Again we see that MLE values demonstrate "implosion" breakdown as before but it stabilizes, and again the L_2E estimates demonstrate robustness. The second observation is that again the price paid by the L_2E for less bias is increased variance (Table 5.2). The L_2E's sample standard error is greater than the MLE's. The third observation is that the L_2E regression coefficients appear to be "exploding" a bit when $w = w_{opt} = 200/(200 + N)$ where N is the number of outliers added.

[1]Note that θ does not include the intercept θ_0.

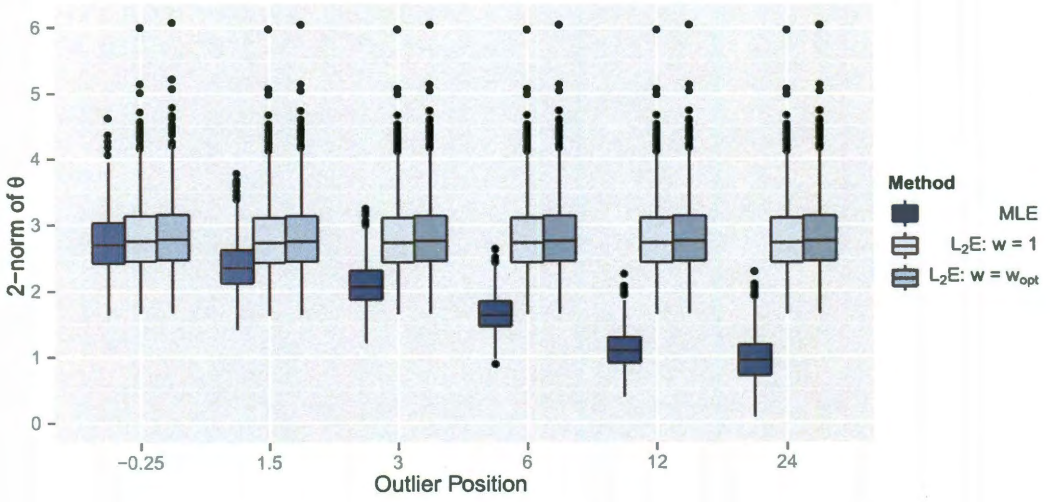


Figure 5.1: The 2-norm of the regression coefficients (intercept not included) as a function of a single outliers position.

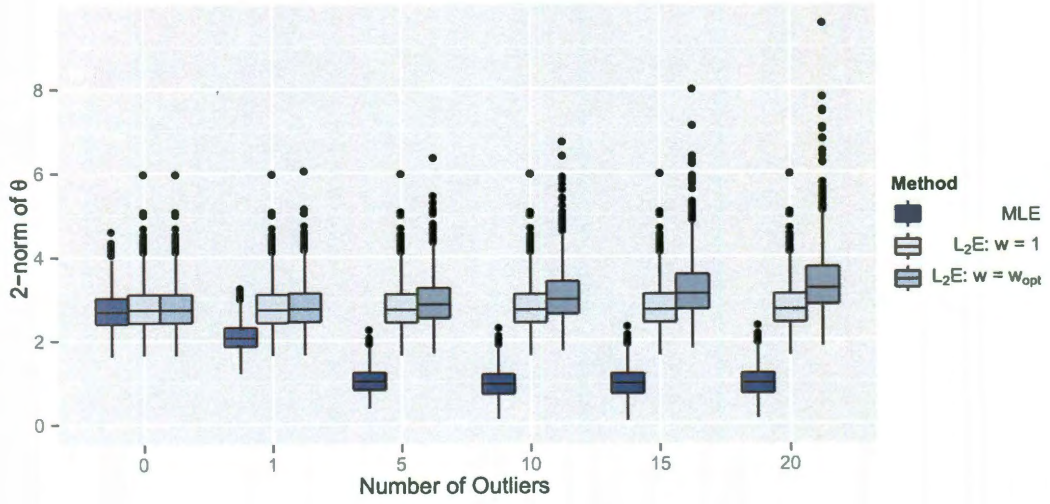


Figure 5.2: The 2-norm of the regression coefficients (intercept not included) as a function of the number of outliers at a fixed position.

Table 5.1: Varying the location of a single outlier: The true parameter value is $\theta = (0, 1, 0.5, 1, 2)^T$. The L_2E calculated θs are essentially unbiased regardless of the location of the outlier. In contrast, the MLE calculated θ become very biased as the outlier position ranges from -0.25 to 24. The unbiasedness of the L_2E comes at a price of increased variance. The sample standard error of the L_2E is greater than that of the MLE for all outlier positions.

		MLE		L_2E (w = 1)		L_2E (w = w_{opt})	
Outlier Position	Coefficient	mean	std	mean	std	mean	std
-0.25	θ_0	-0.002	0.182	-0.005	0.192	-0.005	0.194
	θ_1	1.032	0.434	1.063	0.480	1.073	0.485
	θ_2	0.526	0.424	0.539	0.463	0.543	0.468
	θ_3	1.047	0.439	1.079	0.482	1.088	0.488
	θ_4	2.110	0.487	2.181	0.572	2.200	0.580
1.5	θ_0	-0.024	0.168	0.002	0.192	0.002	0.194
	θ_1	0.868	0.394	1.052	0.476	1.061	0.481
	θ_2	0.401	0.391	0.532	0.460	0.536	0.465
	θ_3	0.880	0.396	1.068	0.478	1.077	0.484
	θ_4	1.860	0.430	2.160	0.567	2.180	0.575
3	θ_0	-0.022	0.157	0.002	0.192	0.002	0.194
	θ_1	0.732	0.368	1.054	0.476	1.063	0.481
	θ_2	0.296	0.369	0.533	0.460	0.537	0.465
	θ_3	0.743	0.368	1.069	0.478	1.078	0.484
	θ_4	1.662	0.392	2.163	0.567	2.182	0.575
6	θ_0	-0.020	0.142	0.002	0.192	0.002	0.194
	θ_1	0.508	0.337	1.054	0.476	1.063	0.481
	θ_2	0.112	0.344	0.533	0.460	0.537	0.465
	θ_3	0.516	0.334	1.069	0.478	1.078	0.484
	θ_4	1.350	0.347	2.163	0.567	2.183	0.575
12	θ_0	-0.018	0.128	0.002	0.192	0.002	0.194
	θ_1	0.153	0.325	1.054	0.476	1.063	0.481
	θ_2	-0.201	0.336	0.533	0.460	0.537	0.465
	θ_3	0.158	0.316	1.069	0.478	1.078	0.484
	θ_4	0.906	0.317	2.163	0.567	2.183	0.575
24	θ_0	-0.011	0.124	0.002	0.192	0.002	0.194
	θ_1	-0.088	0.330	1.054	0.476	1.063	0.481
	θ_2	-0.431	0.331	0.533	0.460	0.537	0.465
	θ_3	-0.086	0.315	1.069	0.478	1.078	0.484
	θ_4	0.641	0.324	2.163	0.567	2.183	0.575

Table 5.2: Varying the number of outliers at a fixed location: The true parameter value is $\theta = (0, 1, 0.5, 1, 2)^T$. The L_2E calculated θs are essentially unbiased regardless of the number of outliers. In contrast, the MLE calculated θ become very biased as outliers are added. The unbiasedness of the L_2E comes at a price of increased variance. The sample standard error of the L_2E is greater than that of the MLE for all numbers of outliers.

		MLE		L_2E (w = 1)		L_2E (w = w_{opt})	
Number of Outliers	Coefficient	mean	std	mean	std	mean	std
0	θ_0	0.0049	0.1824	0.0021	0.1923	0.0021	0.1923
	θ_1	1.0258	0.4326	1.0537	0.4759	1.0537	0.4759
	θ_2	0.5213	0.4225	0.5327	0.4599	0.5327	0.4599
	θ_3	1.0405	0.4376	1.0690	0.4782	1.0690	0.4782
	θ_4	2.0994	0.4853	2.1630	0.5666	2.1630	0.5666
1	θ_0	-0.0221	0.1573	0.0021	0.1923	0.0021	0.1943
	θ_1	0.7324	0.3679	1.0537	0.4759	1.0629	0.4814
	θ_2	0.2956	0.3690	0.5327	0.4599	0.5371	0.4648
	θ_3	0.7431	0.3681	1.0690	0.4782	1.0784	0.4838
	θ_4	1.6620	0.3924	2.1629	0.5666	2.1825	0.5748
5	θ_0	-0.0898	0.1258	0.0021	0.1923	0.0020	0.2026
	θ_1	0.0864	0.3201	1.0537	0.4759	1.1008	0.5046
	θ_2	-0.2628	0.3272	0.5327	0.4599	0.5554	0.4854
	θ_3	0.0905	0.3082	1.0690	0.4782	1.1166	0.5071
	θ_4	0.8300	0.3125	2.1629	0.5666	2.2625	0.6095
10	θ_0	-0.1101	0.1237	0.0021	0.1923	0.0022	0.2148
	θ_1	-0.0735	0.3296	1.0536	0.4759	1.1511	0.5403
	θ_2	-0.4167	0.3332	0.5326	0.4599	0.5791	0.5132
	θ_3	-0.0709	0.3153	1.0690	0.4782	1.1666	0.5391
	θ_4	0.6586	0.3226	2.1628	0.5667	2.3673	0.6600
15	θ_0	-0.1172	0.1237	0.0021	0.1923	0.0021	0.2274
	θ_1	-0.1268	0.3354	1.0536	0.4759	1.2033	0.5784
	θ_2	-0.4696	0.3385	0.5326	0.4599	0.6044	0.5437
	θ_3	-0.1245	0.3208	1.0689	0.4782	1.2188	0.5746
	θ_4	0.6048	0.3282	2.1627	0.5667	2.4771	0.7188
20	θ_0	-0.1216	0.1238	0.0021	0.1923	0.0018	0.2401
	θ_1	-0.1586	0.3393	1.0535	0.4759	1.2577	0.6182
	θ_2	-0.5016	0.3423	0.5326	0.4599	0.6313	0.5779
	θ_3	-0.1566	0.3246	1.0689	0.4782	1.2738	0.6159
	θ_4	0.5735	0.3318	2.1626	0.5668	2.5923	0.7858

5.3 Variable Selection in High Dimensions

In the variable selection experiment we considered a high dimensional variation on the first scenario. We generated 1000 data sets each with $n = 450$ using the model given in equation (2.3) where $\theta \in \mathbb{R}^{20,000}$ whose first 30 components were 1 and whose other 19,970 components were 0. The covariates $\mathbf{x}_i$ again were drawn from one of two populations. For all i, $x_{i0} = 1$ and ϵ_{ij} were iid $N(0,1)$ for all i and j. For $i = 1, \ldots, 200$, $x_{ij} = 0.3 + 0.75\epsilon_{ij}$ for $j = 1, \ldots, 30$. For $i = 201, \ldots, 400$, $x_{ij} = -0.3 + 0.75\epsilon_{ij}$ for $j = 1, \ldots, 30$. We then added 50 inliers, $\{(y_i, \mathbf{x}_i)\}$ where $y_i = 0$ and $x_{ij} = 0.1 + 0.25\epsilon_{ij}$ for $i = 401, \ldots, 450$ and $j = 1, \ldots, 30$. For $j = 31, \ldots, 20,000$ and $i = 1, \ldots, 450$, $x_{ij} = 0.75\epsilon_{ij}$.

We then performed Elastic Net penalized regression with the MLE and L_2E. To perform model selection we generated regularization paths. That is we calculated penalized regression coefficients for a range of λ values. We then compared the fits for different λ values using the robust BIC criterion described in Chapter 3. The model with the lowest estimated prediction error was selected. To perform the elastic net penalized logistic regression we used the **glmnet** package in R [22].

Let $\hat{\theta}_j^{(k)}$ denote the regression coefficient for the jth covariate in the kth replicate. Then the expected number of true positives was estimated by

$$\frac{1}{1000} \sum_{k=1}^{1000} \sum_{j=1}^{30} 1(\hat{\theta}_j^{(k)} \neq 0),$$

and similarly the expected number of false positives was estimated by

$$\frac{1}{1000} \sum_{k=1}^{1000} \sum_{j=31}^{20000} 1(\hat{\theta}_j^{(k)} \neq 0).$$

Figure 5.3 and Figure 5.4 show the number of true positives and false positives respectively for each method. We see that both methods selected at least one True Positive,

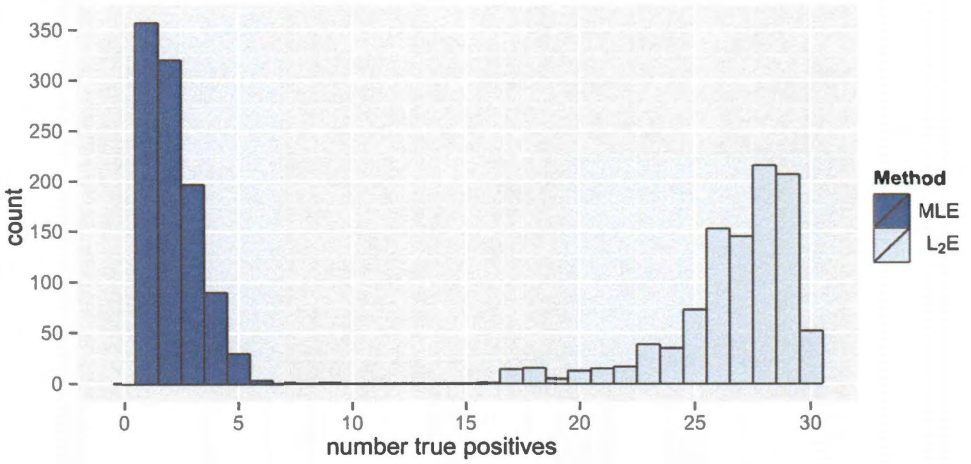


Figure 5.3: Comparison of true positives selected by MLE and L_2E

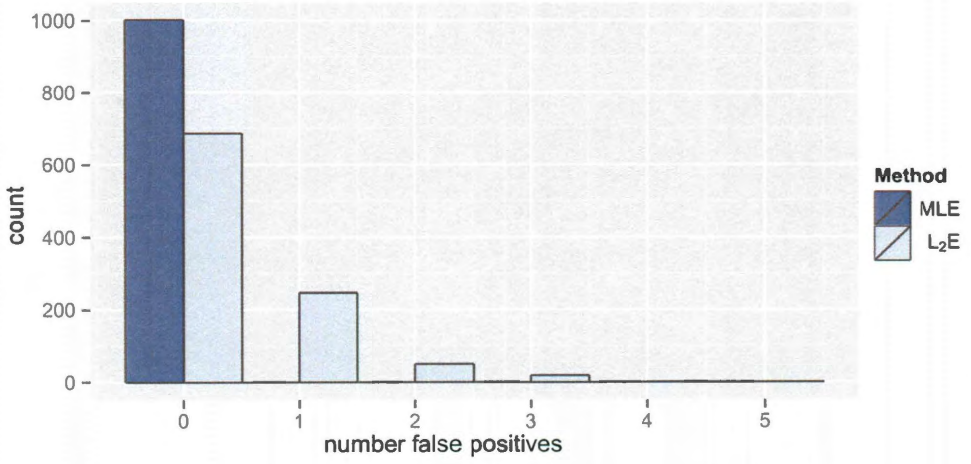


Figure 5.4: Comparison of false positives selected by MLE and L_2E

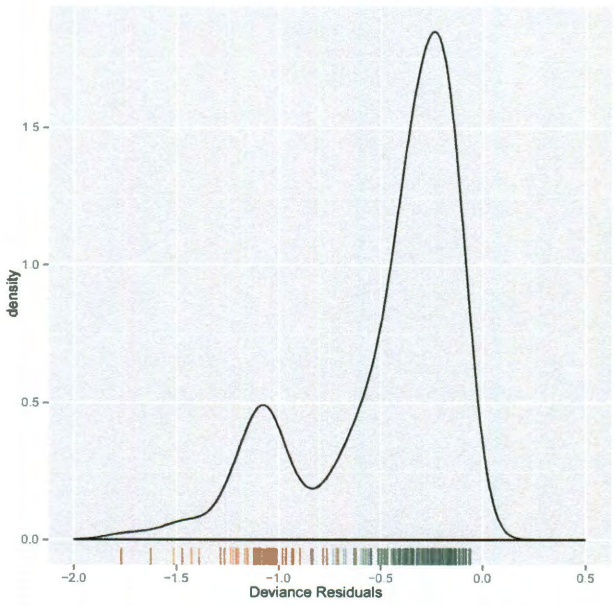


Figure 5.5: Density estimate of deviance residuals that have $y_i = 0$ (L_2E)

however L_2E did substantially better on average. This is consistent with the observed behavior in the estimation experiments. The presence of outliers causes *implosion* breakdown in the MLE; the estimated regression coefficients are biased towards zero. The ℓ_1 penalty also biases estimates toward zero. Thus, the presence of outliers can drive the estimated effect of true covariates sufficiently close to zero that the ℓ_1 penalty eliminates them. We also see that while the MLE did not incur any false positives, out of 1000 replicates the L_2E selected at most 5 false positives. Thus, the L_2E's superior sensitivity in the presence of inliers does not come at a cost of decreased specificity.

We take a closer look at one of the replicates to demonstrate how outliers and inliers can be detected. We simply compute the deviance residuals. Figure 5.5 shows a rug and density estimate of the deviance residuals for the label with inlying observations, i.e. $y_i = 0$ using estimates generated by L_2E. Figure 5.6 shows the same information as generated by MLE. As expected the robust procedures produce fits in which the inliers have large deviances.

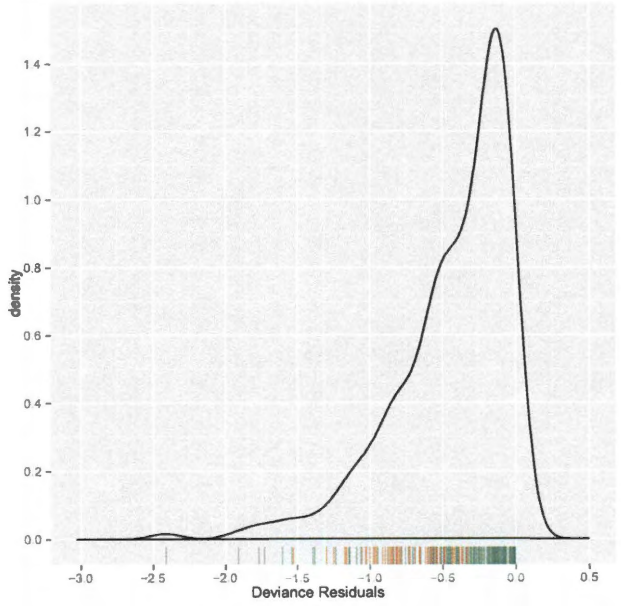


Figure 5.6: Density estimate of deviance residuals for observations that have $y_i = 0$ (MLE)

CHAPTER 6

REAL DATA

6.1 Galaxy Data

The European Southern Observatory (ESO) used the Wide-Field Imager at the MPG/ESO 2.2-m telescope at La Silla observatory, Chile, to image a region in the sky known as the Chandra Deep Field South (CDFS) to catalog the celestial bodies in it [56].

The ESO has made their catalog of 3,438 galaxies in the CDFS publicly available. One quantity of interest is the mean red-shift (Mcz) of a galaxy; this is a surrogate measure for the speed at which a galaxy is moving away from ours. Additionally, for each galaxy information about the absolute magnitudes in 10 frequency bands is available in addition to flux measurements in 13 bands ranging from 420 nm (ultraviolet) to 915 nm (far red). Figure 6.1 shows a histogram of Mcz which shows clear bimodality. To test the performance of the logistic L_2E we create threshold the Mcz of galaxies to create binary response variables by assigning a response $y_i = 1$ if the ith galaxy's Mcz exceeds 0.45 and 0 otherwise. We are interested in modeling the binary measures of Mcz with the other 23 bands.

We first inspect the variables. Figures 6.2, 6.3 and 6.4 show three matrices of smoothed scatter plots between pairs of the 23 bands and Mcz. Figure 6.4 compares 5 magnitude bands with 5 flux bands. The combinations not shown were very similar. Variables are

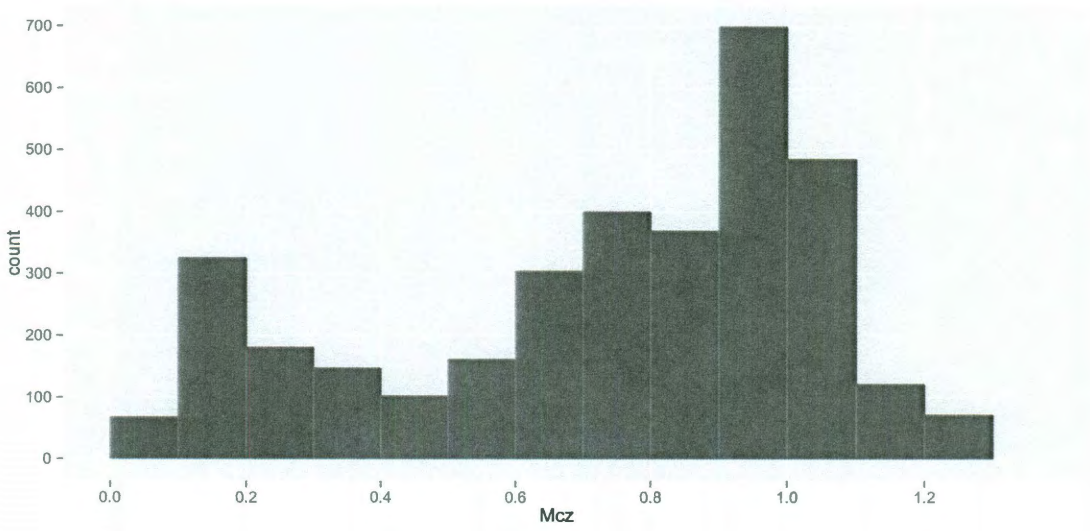


Figure 6.1: Distribution of red shifts (Mcz) of galaxies in the CDFS

ordered from left to right as Mcz, the 10 absolute magnitude measurements, followed by the 13 flux measurements. Along the diagonal are univariate density estimates of the variables. Panels above the diagonal contain 2-dimensional histograms of pairs of variables using hexagonal bins; counts are encoded in gray scale. Panels below the diagonal contain the 2-dimensional histograms overlaid with a loess curve shown in red.

We make three observations. First, the flux measurements are all positively correlated with each other but poorly correlated with either the absolute magnitude measurements or Mcz. The absolute magnitude measurements are also positively correlated with each other, and negatively correlated with Mcz. The high correlation between the covariates suggests including a ridge penalty. The fact that just over half the predictors are poorly correlated with Mcz suggests a LASSO penalty to induce sparsity. So, even though this data set is in the regime of $n > p$, it is still useful to perform Elastic Net penalized regression.

The third observation is that the loess smooth in the third column of the last row, the scatter between Mcz and BjMAG, is different from all other loess smooths between Mcz

and the other MAG measurements. Indeed there is a clear outlier in BjMAG. Observation number 2 has a BjMAG of 17.86 whereas all other observations have BjMAG that range between -23.15 and -7.58. Most likely observation 2 had a BjMAG of -17.86. We will proceed with comparing the MLE and logistic L_2E without making any corrections first to see the effects of this outlier. Later we will make the correction and compare the MLE and logistic L_2E.

We performed the two-step procedure described in Section 3.6 using the robust BIC for both the MLE and L_2E using a fixed parameter $\alpha = 0.05$ given the high degree of collinearity among the predictors. Logistic L_2E selected the 10 magnitude variables (UjMAG, BjMAG, VjMAG, usMAG, gsMAG, rsMAG, UbMAG, BbMAG, VnMAG, and S280MAG). The MLE selected 8 magnitude variables (UjMAG, VjMAG, usMAG, gsMAG, UbMAG, BbMAG, VnMAG, S280MAG). Notably the MLE did not include BjMAG.

We plotted the fitted probabilities against the Mcz values in Figure 6.5(a) and Figure 6.5(b). In both the panels are divided into four quadrants. The points to the left of the division have Mcz less than or equal to 0.45 and those on the right greater than 0.45. Points above the horizontal division have fitted probabilities of at least 0.5 and those below less than 0.5. Thus, the upper left and lower right quadrants denote points which have good agreement between their true underlying Mcz and their fitted probabilities. We see that both the MLE and logistic L_2E find fits in which many low Mcz points have high probabilities of having a class label of 1. Nonetheless there are important differences.

The most notable difference is between the number of points in the lower right hand corner corresponding to high Mcz points with lower fitted probabilities of being labelled a 1. Note in particular that the MLE does not detect the outlier (in blue) and places it in the lower right hand quadrant because BjMAG was not selected by the MLE whereas the logistic L_2E does. As we saw in the simulated data examples outliers can cause missed detections in variable selection. We plotted the 10 chosen covariate values (jittered) strat-

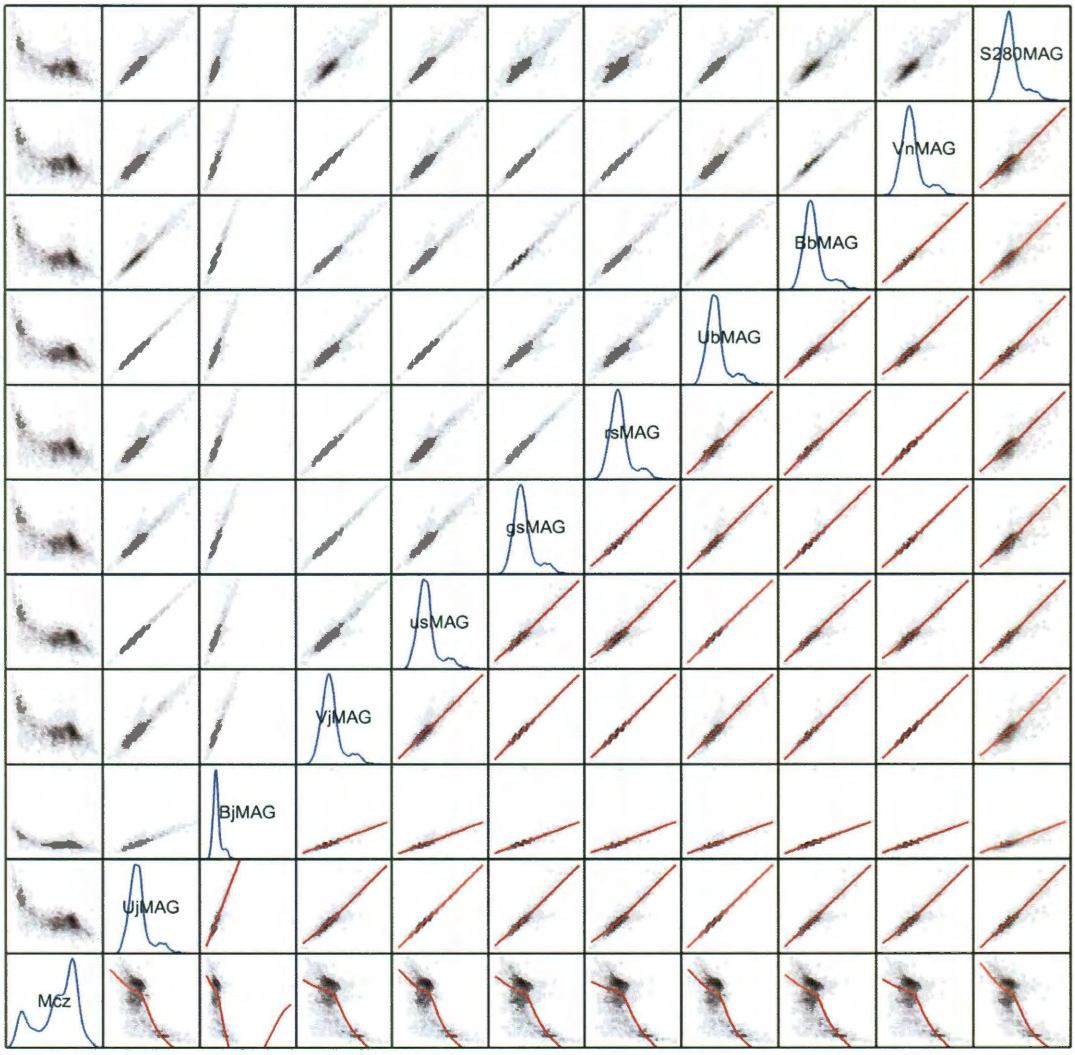


Figure 6.2: Scatter Plot Matrix of Mcz and intensity bands.

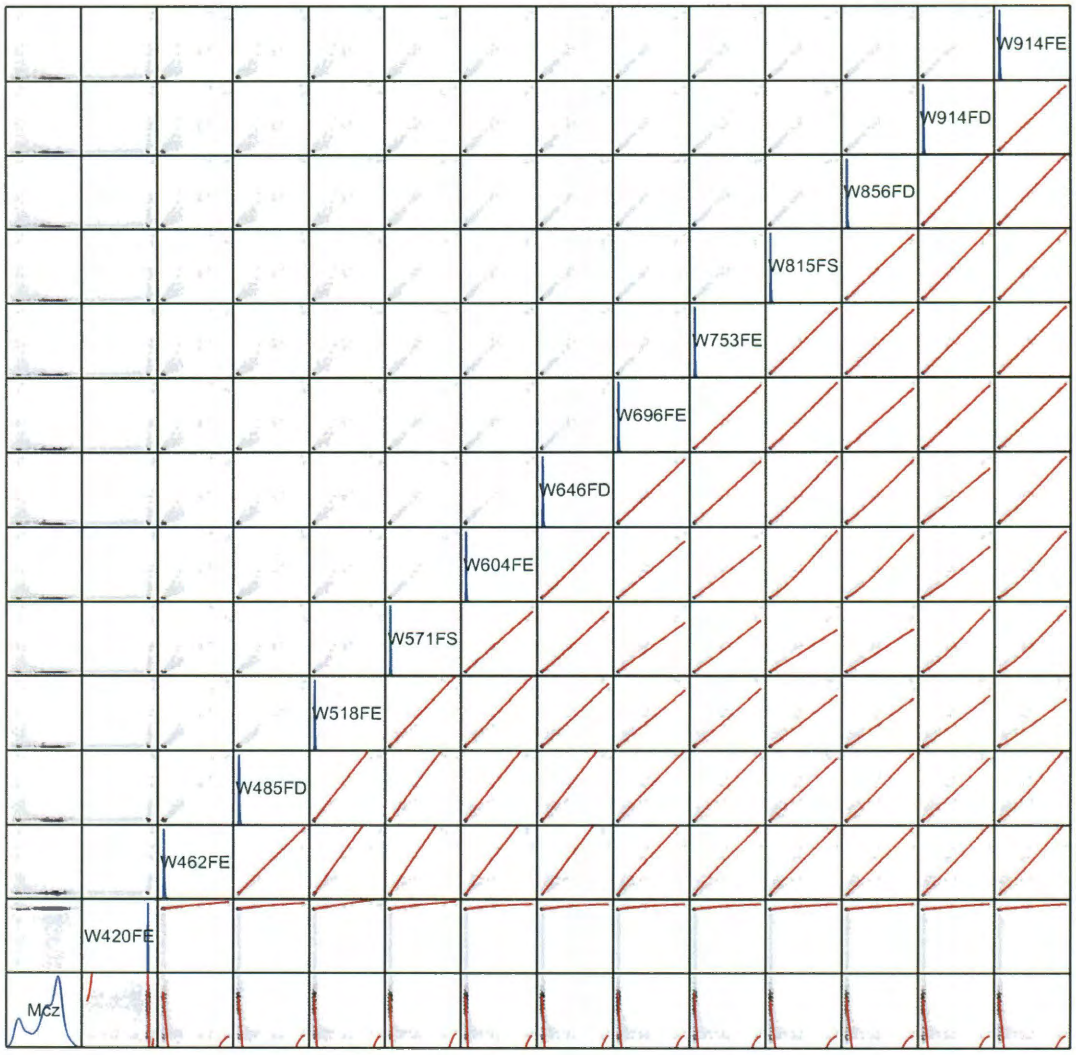


Figure 6.3: Scatter Plot Matrix of Mcz and flux bands.

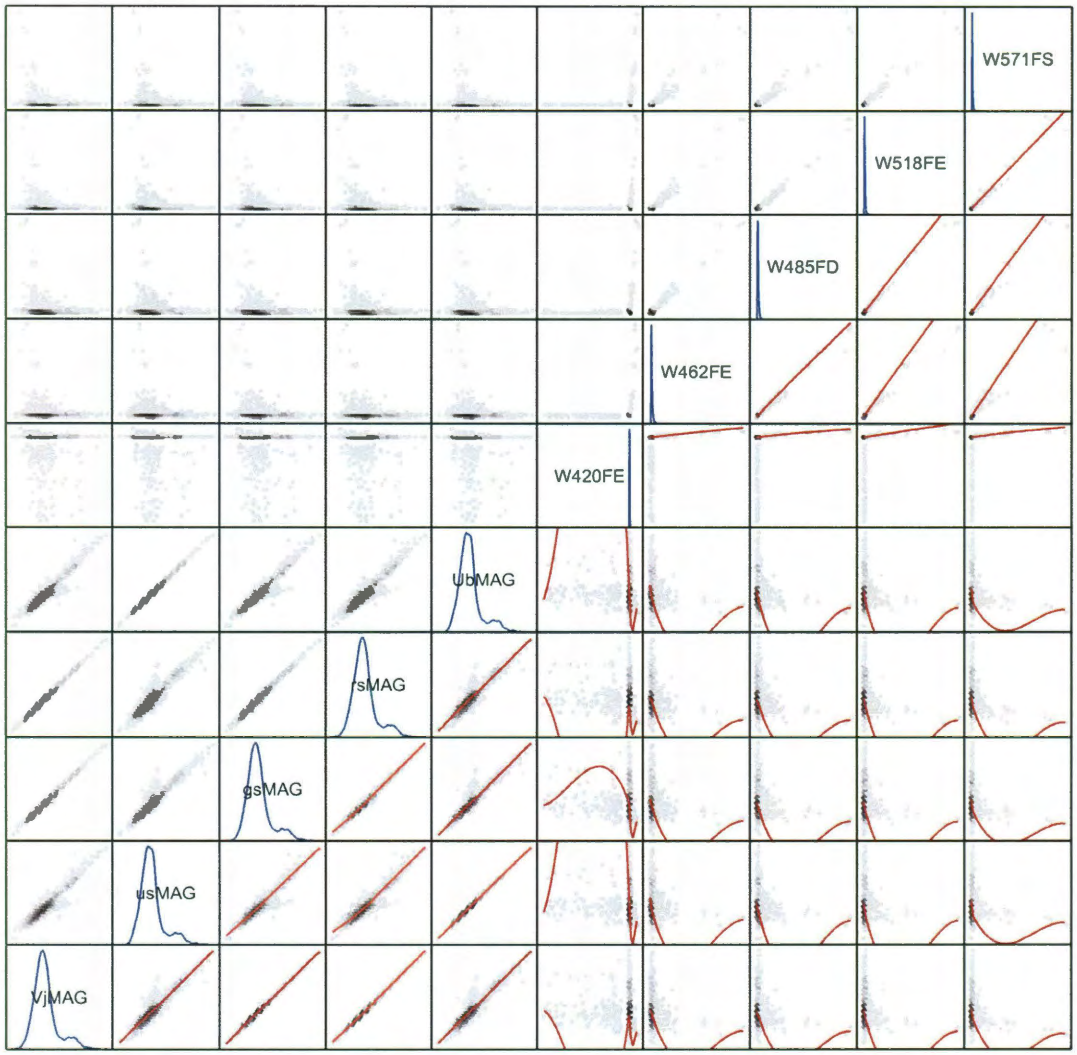
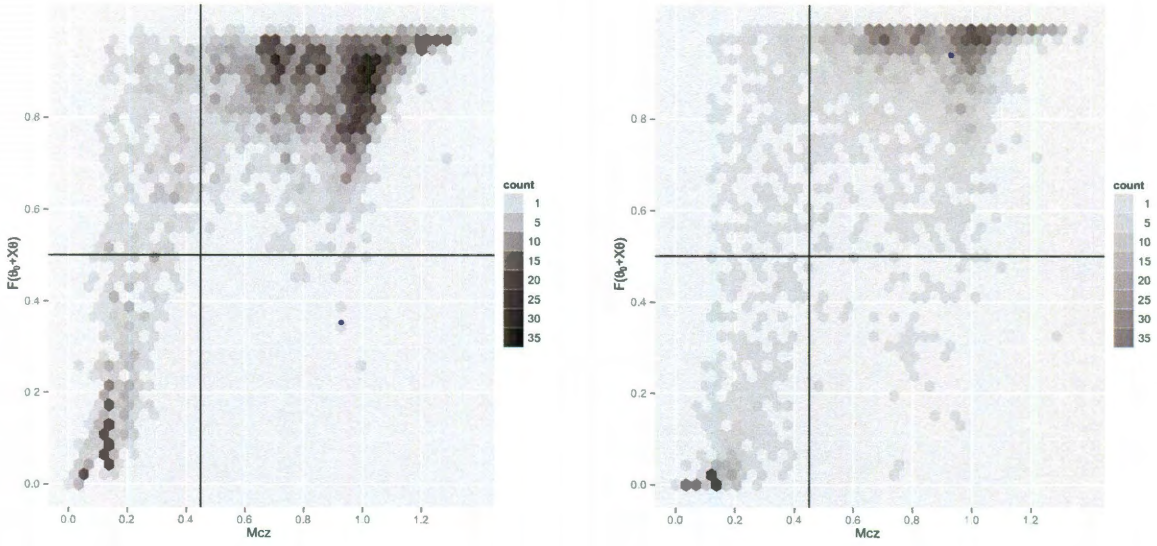


Figure 6.4: Scatter Plot Matrix of 5 intensity bands and 5 flux bands.



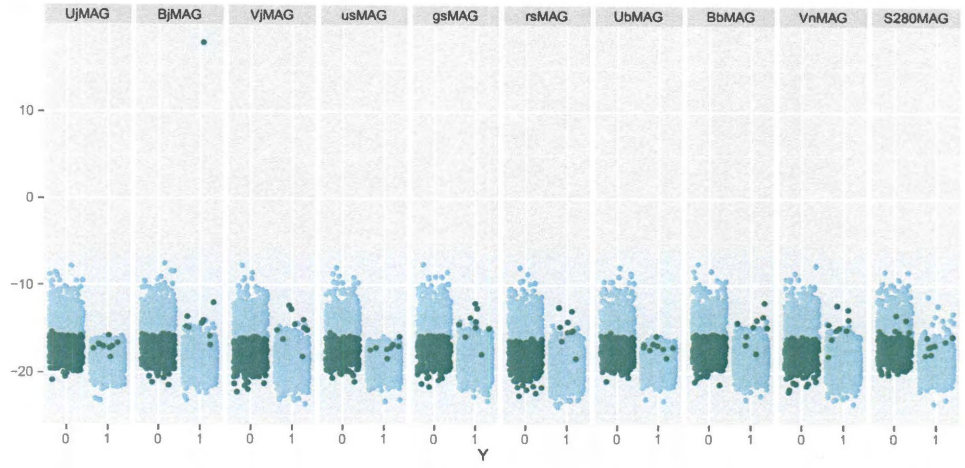
(a) L_2E (b) MLE

Figure 6.5: Fitted probabilities versus Mcz. Blue dot denotes transcription outlier

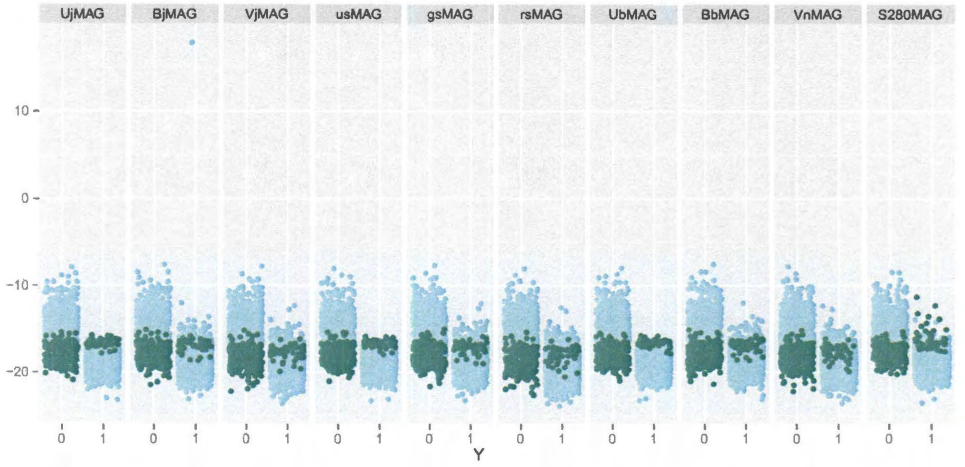
ified by their class labels in Figure 6.6(a); dark green points correspond to points whose fitted probabilities disagreed with their observed class label, i.e. $P(Y_i = 1|\mathbf{x}_i) < 0.5$ and $P(Y_i = 0|\mathbf{x}_i) > 0.5$. Note a threshold other than 0.5 could be used. For convenience we refer to these points as "misclassified." As would be hoped the outlier is among the misclassified points in the logistic L_2E regression. Also the misclassified points are relatively extreme in each of the 10 covariate spaces.

Figure 6.6(b) plot the values of the same 10 covariates although only 8 were chosen; analogously dark green points correspond to points misclassified by the MLE.

We then repeated the above procedure after correcting the sign of the outlying observation. The logistic L_2E selected the same 10 covariates. The MLE selected a set of 7 covariates (UjMAG, BjMAG, usMAG, gsMAG, UbMAG, BbMAG, S280MAG). Interestingly after correcting for the transcription error in it, the MLE selected BjMAG. Figure 6.7(a) and Figure 6.7(b) show a scatter of fitted probabilities against Mcz. The logistic L_2E is virtually unchanged from before except for the former outlier. The misclassified points

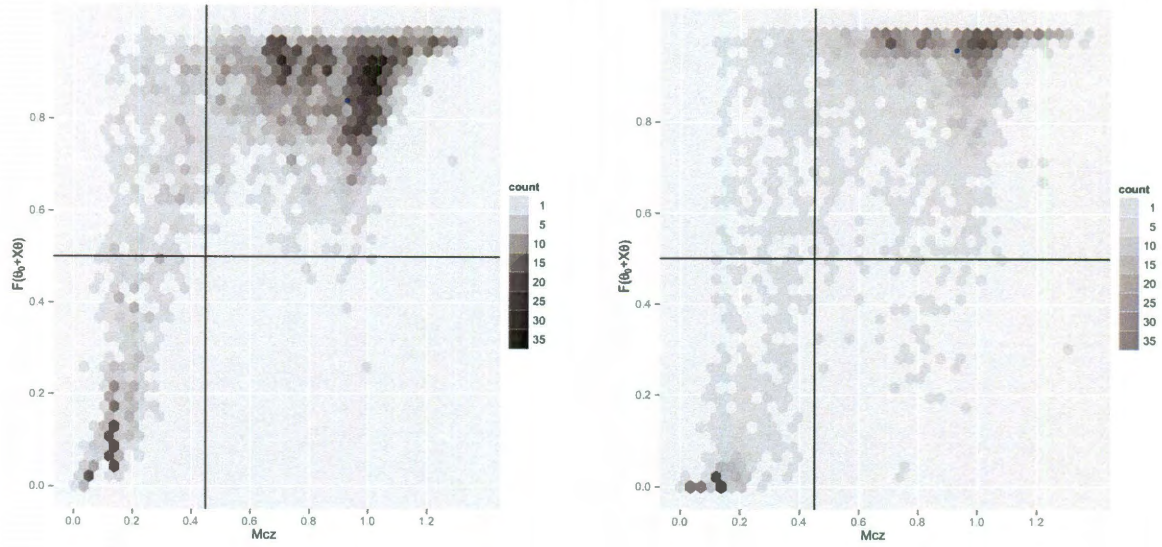


(a) L_2E: All ten covariates shown were selected.



(b) MLE: UjMAG, VjMAG, usMAG, gsMAG, UbMAG, BbMAG, VnMAG, and S280MAG selected.

Figure 6.6: Covariate values for magnitude bands. Dark green points correspond to points in the upper left hand and lower right hand quadrants in Figures 6.5(a) and 6.5(b).

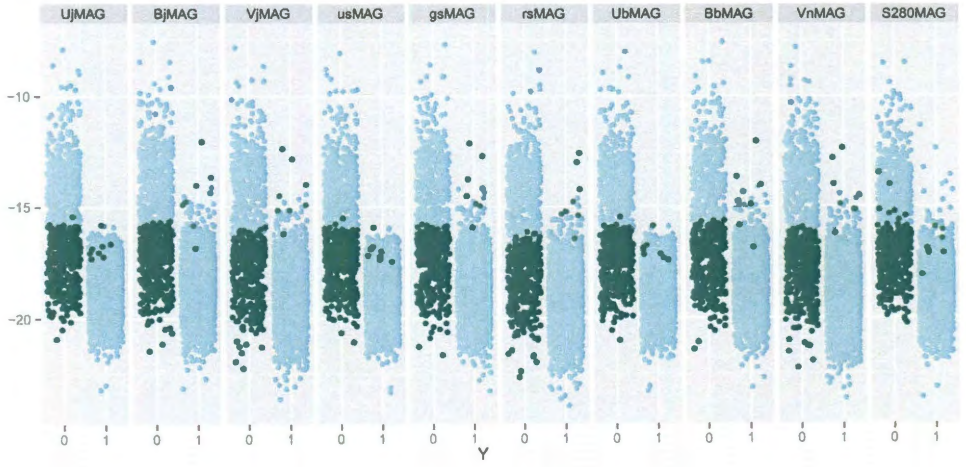


(a) L_2E (b) MLE

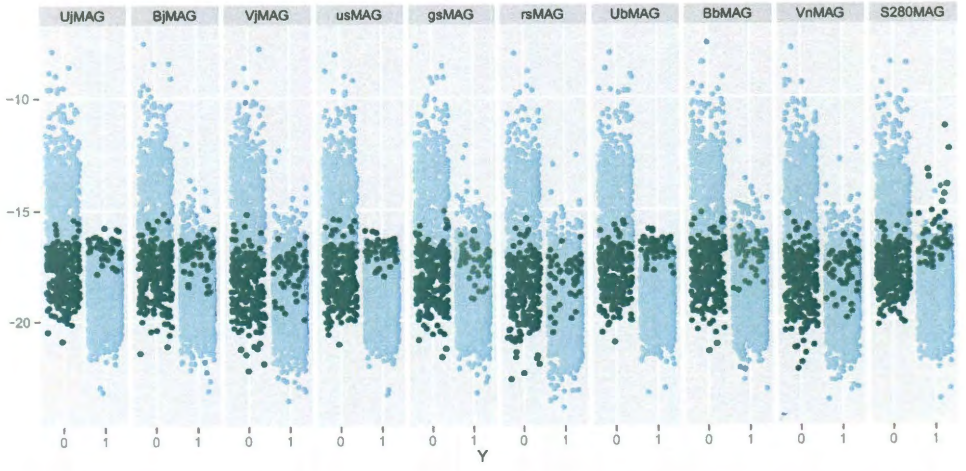
Figure 6.7: Fitted probabilities versus Mcz after correcting transcription outlier. Blue dot denotes point with former transcription error.

do vary somewhat for the MLE after correcting the outlier as seen in Figure 6.8(a) and Figure 6.8(b).

Finally, in Figure 6.9(a) and Figure 6.9(b) we color points in Figure 6.7(b) and Figure 6.7(b) that are misclassified by the other method. The question we ask is whether points misclassified by the MLE are also likely to be misclassified by the L_2E. Surprisingly we see that there is little agreement although most points misclassified as a 0 by the MLE will also be misclassified as a 0 by the logistic L_2E. The two methods are finding different explanatory structure in the covariate space. Nonetheless from a prediction point of view the MLE does provide a decent model, but it is likely due to the fact we have so many observations.

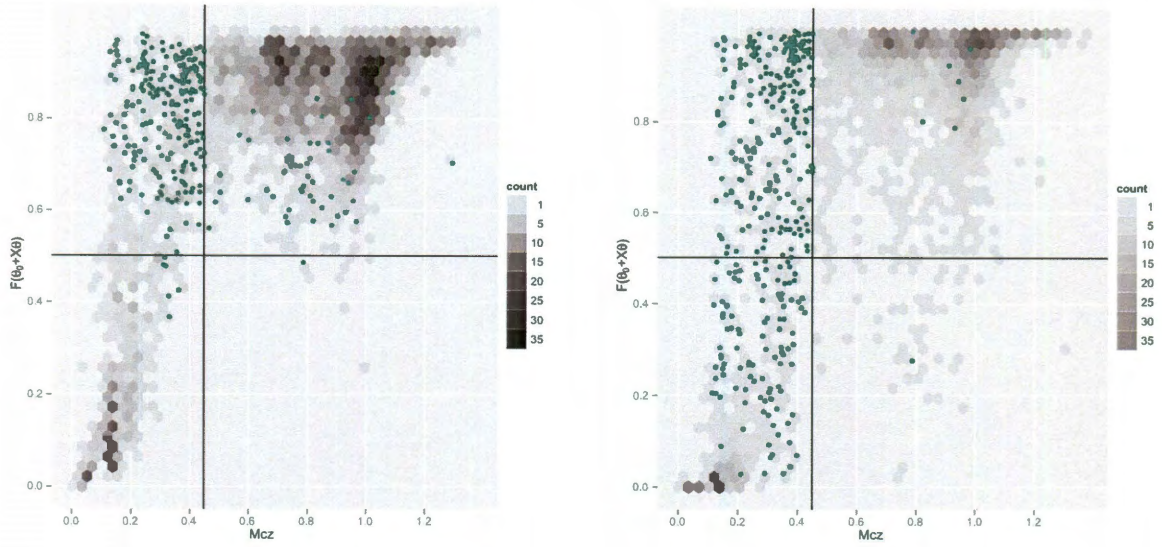


(a) L_2E: All ten covariates shown were selected.



(b) MLE: UjMAG, BjMAG, usMAG, gsMAG, UbMAG, BbMAG, and S280MAG selected.

Figure 6.8: Covariate values for magnitude bands after correcting transcription error. Dark green points correspond to points in the upper left hand and lower right hand quadrants in Figures 6.7(a) and 6.7(b).



(a) L_2E (b) MLE

Figure 6.9: Fitted probabilities versus Mcz. Green points denote points that are misclassified by the other estimation procedure.

6.2 Genome Wide Association Data

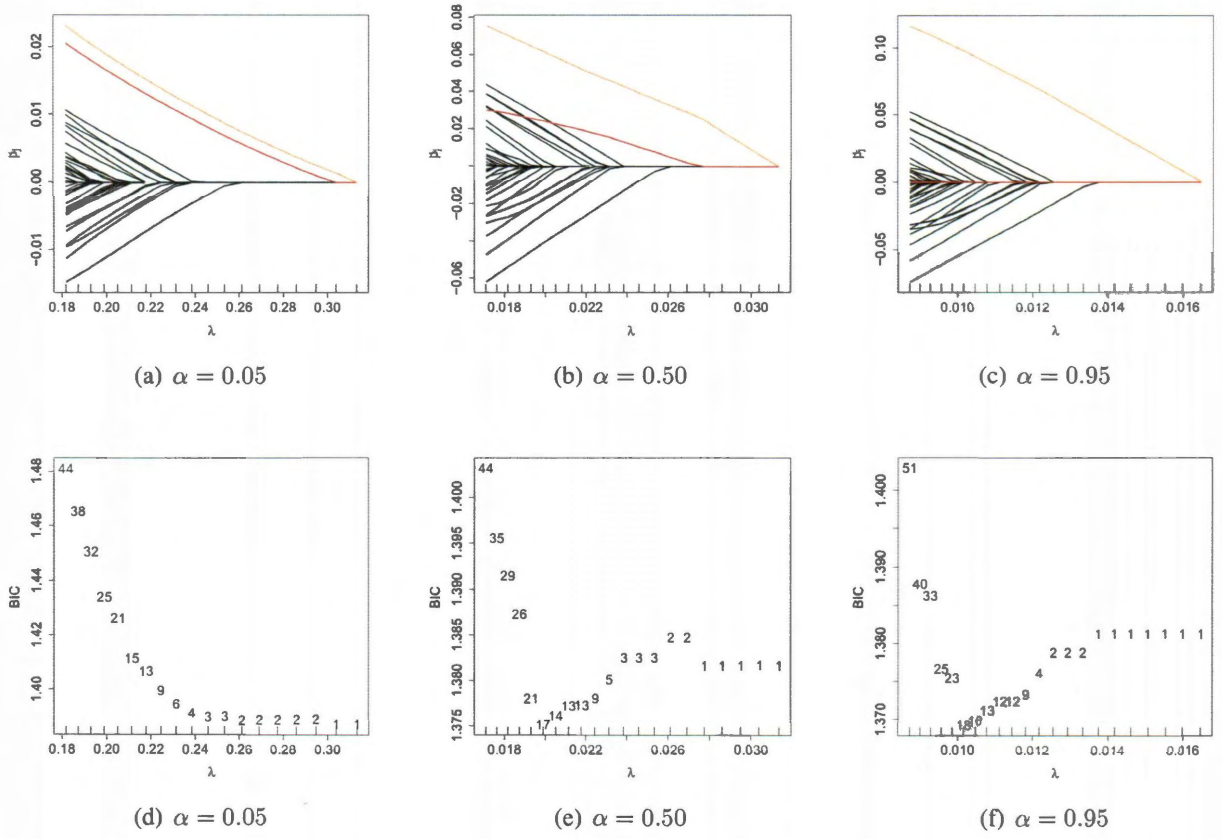
We examine the lung cancer data of Amos et al. [2]. The purpose of this genome wide association study was to identify risk variants for lung cancer. The authors employed a two stage study using 315,450 tagging SNPs in 1,154 current and former (ever) smokers of European ancestry and 1,137 frequency matched, ever-smoking controls from Houston Texas in the discovery stage. The most significant SNPs found in the discovery phases were then tested in a larger replication set. Two SNPs, rs1051730 and rs8034191, on chromosome 15 were found to be significantly associated with lung cancer risk in the validation set.

In this section we reexamine the discovery data using logistic L_2E and the logistic MLE. Note it is current practice of geneticists to do univariate inference with an adjustment for multiple testing and this approach was taken [2]. Taking a multivariate approach as will be done in this section however allows the analyst to take into account dependencies between the SNPs. We begin relatively small by considering only SNPs found on chromosome 15.

We impute missing genotypes at a SNP by using the MACH 1.0 package, a Markov Chain based haplotyper [35]. After missing data is imputed and keeping only imputations with a quality score of at least 0.9, 8,701 SNPs are retained on 1152 cases and 1136 controls.

We performed the two-step procedure described in Section 3.6 using the robust BIC for both the MLE and L_2E using parameter values $\alpha = 0.05, 0.5,$ and 0.95. Here we use the trimmed mean log-deviance in the robust BIC calculation trimming the most extreme 1% measurements in both tails. SNP markers can have a high degree of collinearity due to recombination mechanics. SNPs that are physically close to each other tend to be highly correlated and are said to be in linkage dysequilibrium. The pair rs1051730 and rs8034191 for example are in "high" linkage dysequilibrium. Figure 6.2 summarizes the variable selection results for the logistic L_2E. In the top series of plots are shown regularization paths. The orange curve corresponds to the regression coefficient of rs1051730 and the red corresponds to the coefficient for rs8034191. The bottom series of plots show the resulting BIC values as a function of the penalty parameter λ. The plotted numbers indicate the size of the solution active set. The specific covariates in the active set tallied at a specific value of λ in the bottom plot corresponds to the set of non-zero curves in the top plot. The hashmarks in both figures indicate the values of λ that were used in the regularization path. Figure 6.2 summarizes the same information for the MLE.

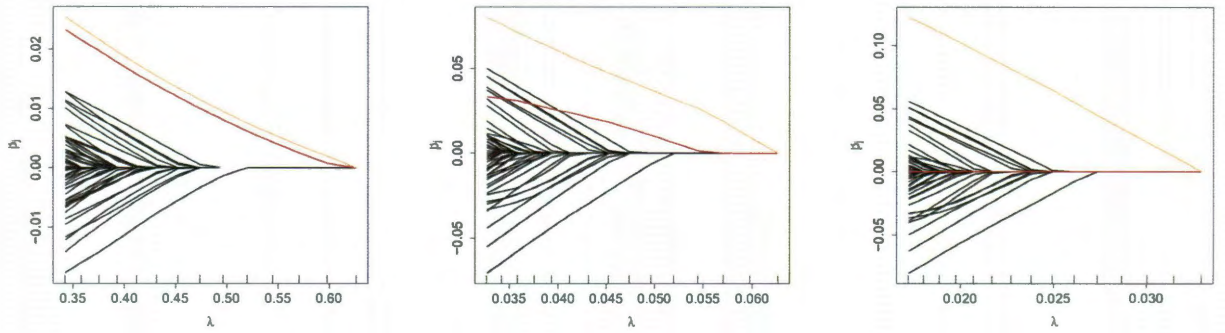
There are three things to note. First, the regularization paths for the L_2E and MLE are almost identical. Second the paths for rs1051730 and rs8034191 behave as would be expected with α. For small α or more ridge like penalty, the two paths become more similar. For large α or more LASSO like penalty, only one of the two correlated predictors enters the model while the other is excluded. The third thing to note is that the BIC curves are different for the two procedures. Even though the same variables are selected for both procedures, it is not surprising that the estimated coefficient values can differ. We see that as a result the logistic L_2E fits tend to select more variables.



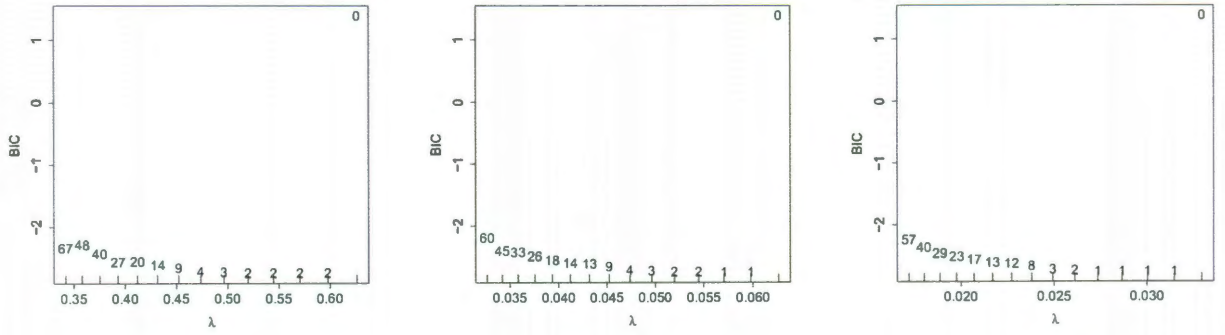
(a) $\alpha = 0.05$ (b) $\alpha = 0.50$ (c) $\alpha = 0.95$

(d) $\alpha = 0.05$ (e) $\alpha = 0.50$ (f) $\alpha = 0.95$

Figure 6.10: Variable Selection of L_2E with BIC



(a) $\alpha = 0.05$ (b) $\alpha = 0.5$ (c) $\alpha = 0.95$



(d) $\alpha = 0.05$ (e) $\alpha = 0.5$ (f) $\alpha = 0.95$

Figure 6.11: Variable Selection of MLE with BIC

CHAPTER 7

DISCUSSION AND CONCLUSIONS

In this final chapter we summarize the contributions of this dissertation and discuss possible extensions and future work.

7.1 Summary of Results

This work has three contributions. Our first contribution is that we have extended the L_2E method to handle binary response variables. Standard logistic regression based on maximum likelihood can suffer implosion breakdown in the presence of outliers. In the classic setting of $n > p$ we have shown that the logistic L_2E provides estimates that are robust against outliers and still provide unbiased estimates when there are no outliers present. The cost is loss in efficiency.

Our second contribution is that we have shown that adding a LASSO-like penalty, the Elastic Net, extends robust regression to robust variable selection and that there are instances where the robust variable selection procedure outperforms standard penalized-likelihood based variable selection procedures. Specifically, our simulation study showed that the penalized MLE can be more conservative and discard more true predictors than the penalized L_2E. The same simulation study also showed that the better sensitivity of the L_2E

did not incur a material loss in specificity. In the context of modern data sets where $n \ll p$, signals are likely to be weak, the logistic L_2E can provide additional power to detect these signals if there are difficult to detect outliers or inliers which will only further weaken said signals.

Thirdly, we have developed a fast algorithm for computing the logistic L_2E solutions using a combination of an MM algorithm and coordinate descent solver. Because the logistic L_2E loss is not convex, our strategy is to solve a series of convex approximations. Specifically, we develop a MM algorithm that uses a quadratic majorization on the logistic L_2E loss exploiting the fact that the loss has bounded curvature. Thus, we convert our non-convex optimization problem into a series of Elastic Net penalized least squares problems. We then apply an established method, coordinate descent, to solve the penalized least squares sub problems very rapidly. We also proved that this algorithm is guaranteed to converge to a stationary point under the regularity condition that the penalized loss has finitely many stationary points.

7.2 Future Work

This dissertation has introduced a framework for robust parametric methods for high dimensional data in the case of binary response variables and demonstrated the potential utility and advantages of this framework as a complement to their parametric penalized-likelihood counterpart. We now discuss extensions that can deepen, improve, and broaden the applicability of these results.

7.3 Applications

LASSO-like penalties perform continuous variable selection and hence open the door to constructing sparse models. The question is whether the sparse set of selected variables is

the "right" sparse set. When prediction error is used to choose the regularization parameter, it has been shown that variable selection with the LASSO is not consistent [34]. That is not to say that variable selection with the LASSO will not be consistent under other circumstances. In fact, consistency can be achieved under a different criterion for choosing the penalty parameter [40]. Establishing the consistency of variable selection for the LASSO, however, comes under rather stringent assumptions on the design matrix $\mathbf{X}$ [40, 60].

On the other hand proving that the LASSO is consistent for prediction error requires no assumptions on $\mathbf{X}$ and relatively mild conditions on the noise [8]. The prediction consistency of the LASSO suggests that while it may select some false positives it selects enough relevant variables to serve as a good variable screening method. If that is the case, then methods like the penalized logistic L_2E that can boost the power of detecting relevant variables in the presence of outliers could have an important role in high-throughput genomics studies. For example, the penalized logistic L_2E could be used in the discovery stage of GWAs.

Another application of the logistic L_2E is in the realm of unsupervised learning of binary data. For example, link relationships in web data can be represented in binary matrices or more generally multiway arrays or tensors [31]. The goal is to find meaningful groupings of the webpages. This can be achieved through generalizations of the SVD and principal components analysis (PCA) for binary data through the machinery of generalized linear models [13]. These generalizations employ an alternating optimization strategy to fit low rank multilinear models of a latent space. A robust version logistic PCA could be achieved by using the minimizing the logistic L_2E as the sub-optimization problem.

7.4 Theory

We demonstrated through simulation examples that the logistic L_2E is robust against varying amounts of outliers. Nonetheless, it would be of interest to characterize more formally

the breakdown behavior of the logistic L_2E. It would also be of interest to work out the asymptotic behavior of the logistic L_2E. It is not uncommon in modern data sets to have many observations even if $n \ll p$. For example in GWAs, there are typically thousands of cases and controls. When we apply a two-stage fitting procedure, the active set may be sufficiently reduced to make $n > p$ for estimation in the second stage. Thus, asymptotic behavior may actually provide a useful description of the logistic L_2E in this second stage.

Finally, while we established that our algorithm will eventually converge to a stationary point, we did not establish the rate at which it does so. Under smoothness assumptions MM algorithms can be shown to have linear local convergence rates [33]. The proofs depend on the implicit function theorem and one strategy for an extension for locally Lipschitz functions would be to use a generalization of the implicit function theorem for locally Lipschitz functions [12].

7.5 Generalizations

The model used in this dissertation can be generalized with respect to the loss and penalty. The loss function could be modified to handle data on other scales, e.g. multinomial logistic L_2E. Scott for example has demonstrated the utility of the L_2E for estimating Gaussian mixture models in low dimensional data [46]. The algorithm for minimizing the penalized L_2E loss could be adapted to fitting Gaussian mixture models in high dimensional data. The challenge would be finding an appropriate majorization to expedite fitting.

In this dissertation we looked exclusively at the Elastic Net because its mixture of ridge and LASSO behavior is appropriate for high dimensional data in which covariates are correlated. Another penalty that should have similar behavior to the Elastic Net is the "Berhu" penalty [41]. Given the importance of computational speed, it would be worthwhile to compare these two similar penalties to see if one works better in practice.

The LASSO has spawned many variants that account for special structure, and the reg-

ularized framework described here can be extended analogously. For example, if covariates are categorical a group LASSO penalty may be more appropriate. Note we did not apply the group LASSO for the GWA in this dissertation since we assumed an additive dose model for the SNPs. There are also arguments for concave penalties like the SCAD [23, 19] which could be applied to our model. In the context of convex losses, concave penalties have been shown to require more computational effort. Since the logistic L_2E loss is already not convex, adding a concave penalty may not be that much more of a computational burden while adding the benefits of concave penalties, such as less bias. Indeed, MM algorithms have been used minimize convex loss functions with concave penalties [64, 44]. Since concave penalties are less biased, employing them may even eliminate the need to do the "relaxed" step in the two-stage estimation procedure.

7.6 Computation

The feasibility of our method hinged on the speed of the computations, but as data sets increase in size our simple algorithm may require some modifications. There are two ways to speed up the outer loop convergence in our algorithm. In this dissertation a very simple quadratic majorization was derived that exploited the bounded curvature of the logistic L_2E loss. We could have instead derived a sharp quadratic majorization [16]. The second improvement would be to speed up local convergence with a quasi-Newton extension of MM algorithms [61].

Finally, there has been recent work in showing that it is possible to "discard" some variables in LASSO optimization problems and obtain the same sparse solution had all variables been included in the optimization. The SAFE rules in [18] guarantee that the solution of the optimization problem over the reduced space is the same as the solution over the full space. The swindle in [57] and the STRONG rules in [53] discard variables more aggressively but do not provide guarantees like the SAFE rules. KKT conditions must be

checked for the solution to the reduced problem. If the KKT conditions are not met, the optimization problem is reformulated with fewer variables are discarded and then the modified reduced problem is solved. If the KKT conditions are ever met for a reduced problem the solution to the reduced problem is a solution to the full problem. In the $n \ll p$ scenario the latter two strategies have been shown to provide significant reductions in computation time. There would be two ways to apply these strategies in our framework. The simpler of the two is to apply the discard rules to the penalized least squares subproblem. More significant savings, however, could be achieved by deriving a version of the sequential STRONG rules in [53] for the logistic L_2E loss.

7.7 Concluding Remarks

As high dimensional data, which is non-trivial to visualize, becomes more common, it becomes more important to have a principled approach to dealing with outliers and inliers. The method developed in this thesis facilitates the use of standard parametric models when some but not necessarily all of the data are well described by the model. It represents a reasonable compromise in weakening standard parametric modeling assumptions without totally abandoning them. The accompanying computational algorithm makes it feasible to compute a regularized robust estimate that can complement the estimates obtained by penalized-likelihood procedures. We hope the examples and method developed in this thesis provide a practical starting point and motivation for further study issues of robustness in high dimensional data.

REFERENCES

[1] H. AKAIKE, *Information theory and the maximum likelihood principle*, Akademiai Kiado, 1973. 3.6

[2] C. I. AMOS, X. WU, P. BRODERICK, I. P. GORLOV, J. GU, T. EISEN, Q. DONG, Q. ZHANG, X. GU, J. VIJAYAKRISHNAN, K. SULLIVAN, A. MATAKIDOU, Y. WANG, G. MILLS, K. DOHENY, Y.-Y. TSAI, W. V. CHEN, S. A. SHETE, M. R. SPITZ, AND R. S. HOULSTON, *Genome-wide association scan of tag snps identifies a susceptibility locus for lung cancer at 15q25.1*, Nature Genetics, 40 (2008), pp. 616–622. 6.2

[3] A. BASU, I. R. HARRIS, N. L. HJORT, AND M. C. JONES, *Robust and efficient estimation by minimising a density power divergence*, Biometrika, 85 (1998), pp. 549–559. 2.1

[4] A. BIANCO AND V. YOHAI, *Robust estimation in the logistic regression model*, in Robust Statistics, Data Analysis, and Computer Intensive Methods, Lecture Notes in Statistics, H. Rieder, ed., vol. 109, New York, 1996, Springer-Verlag, pp. 17–34. 2.4.1

[5] D. BÖHNING AND B. G. LINDSAY, *Monotonicity of quadratic-approximation algorithms*, Annals of the Institute of Statistical Mathematics, 40 (1988), pp. 641–663. 10.1007/BF00049423. 3.1

[6] H. D. BONDELL, *Minimum distance estimation for the logistic regression model*, Biometrika, 92 (September 2005), pp. 724–731. 2.4.1

[7] S. BOYD AND L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, 2004. 1.1.3

[8] P. BÜHLMANN AND S. VAN DE GEER, *Statistics for High-Dimensional Data: Methods, Theory And Applications*, Springer, 2011. 7.3

[9] R. J. CARROLL AND S. PEDERSON, *On robustness in the logistic regression model*, Journal of the Royal Statistical Society. Series B (Methodological), 55 (1993), pp. pp. 693–706. 2.4.1

[10] S. S. CHEN, D. L. DONOHO, AND M. A. SAUNDERS, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing, 20 (1998), pp. 33–61. 1

[11] J. F. CLAERBOUT AND F. MUIR, *Robust modeling with erratic data*, Geophysics, 38 (1973), pp. 826–844. 1

[12] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, 1983. 4.1, 4.1, 4.1, 7.4

[13] M. COLLINS, S. DASGUPTA, AND R. SCHAPIRE, *A generalization of principal component analysis to the exponential family*, Advances in neural information processing systems, 1 (2002), pp. 617–624. 7.3

[14] J. B. COPAS, *Binary regression models for contaminated data*, Journal of the Royal Statistical Society. Series B (Methodological), 50 (1988), pp. pp. 225–265. 2.4.1

[15] C. CROUX, C. FLANDRE, AND G. HAESBROECK, *The breakdown behavior of the maximum likelihood estimator in the logistic regression model*, Statistics & Probability Letters, 60 (2002), pp. 377–386. 1.4, 1.4, 1

[16] J. DE LEEUW AND K. LANGE, *Sharp quadratic majorization in one dimension*, Computational Statistics and Data Analysis, 53 (2009), pp. 2471–2484. 7.6

[17] B. EFRON, T. HASTIE, I. JOHNSTONE, AND R. TIBSHIRANI, *Least angle regression*, Annals of Statistics, 32 (2004), pp. 407–499. 3.6

[18] L. EL GHAOUI, V. VIALLON, AND T. RABBANI, *Safe Feature Elimination in Sparse Supervised Learning*, Tech. Report UCB/EECS-2010-126, EECS Department, University of California, Berkeley, Sep 2010. 7.6

[19] J. FAN AND R. LI, *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 96 (2001), pp. pp. 1348–1360. 1.1.3, 7.5

[20] I. E. FRANK AND J. H. FRIEDMAN, *A statistical view of some chemometrics regression tools*, Technometrics, 35 (1993), pp. pp. 109–135. 1.1.3

[21] J. FRIEDMAN, T. HASTIE, H. HÖFLING, AND R. TIBSHIRANI, *Pathwise coordinate optimization*, Annals of Applied Statistics, 1 (2007), pp. 302–332. 3, 3.8

[22] J. H. FRIEDMAN, T. HASTIE, AND R. TIBSHIRANI, *Regularization paths for generalized linear models via coordinate descent*, Journal of Statistical Software, 33 (2010), pp. 1–22. 1.3, 5.3

[23] W. FU AND K. KNIGHT, *Asymptotics for lasso-type estimators*, Annals of Statistics, 28 (2000), pp. 1356–1378. 1.1.3, 7.5

[24] A. GENKIN, D. D. LEWIS, AND D. MADIGAN, *Large-scale bayesian logistic regression for text categorization*, Technometrics, 49 (2007), pp. 291–304. 1.3

[25] K. R. HESS, K. ANDERSON, W. F. SYMMANS, V. VALERO, N. IBRAHIM, J. A. MEJIA, D. BOOSER, R. L. THERIAULT, A. U. BUZDAR, P. J. DEMPSEY, R. ROUZIER, N. SNEIGE, J. S. ROSS, T. VIDAURRE, H. L. GÓMEZ, G. N. HORTOBAGYI, AND L. PUSZTAI, *Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer*, Journal of Clinical Oncology, 24 (2006), pp. 4236–4244. 1

[26] A. E. HOERL AND R. W. KENNARD, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, 12 (1970), pp. pp. 55–67. 1.1.3

[27] D. HUNTER AND K. LANGE, *A Tutorial on MM Algorithms.*, The American Statistician, 58 (2004), pp. 30–38. 3

[28] N. L. JOHNSON AND B. L. WELCH, *Applications of the non-central t-distribution*, Biometrika, 31 (1940), pp. pp. 362–389. 1.2

[29] J. KIM AND C. SCOTT, *Performance analysis for L_2 kernel classification*, in Advances in Neural Information Processing Systems 21 (NIPS 2008), D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds., 2009, pp. 833–840. 2.4.1

[30] J. KIM AND C. SCOTT, *L_2 kernel classification*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32 (2010), pp. 1822 –1831. 2.4.1

[31] T. KOLDA AND B. BADER, *The TOPHITS model for higher-order web link analysis*, in Proceedings of the SIAM Data Mining Conference Workshop on Link Analysis, Counterterrorism and Security, 2006. 7.3

[32] H. R. KÜNSCH, L. A. STEFANSKI, AND R. J. CARROLL, *Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models*, Journal of the American Statistical Association, 84 (1989), pp. pp. 460–466. 2.4.1

[33] K. LANGE, *Numerical Analysis for Statisticians*, Springer, 2010. 4.2, 4.2, 4.2, 7.4

[34] C. LENG, Y. LIN, AND G. WAHBA, *A note on the Lasso and related procedures in model selection*, Statistica Sinica, 16 (2006), pp. 1273–1284. 7.3

[35] Y. LI, J. DING, AND G. R. ABECASIS, *Mach 1.0: rapid haplotype reconstruction and missing genotype inference.*, American Journal of Human Genetics, S79 (2006), p. 2290. 6.2

[36] Z. LIU, F. JIANG, G. TIAN, S. WANG, F. SATO, S. J. MELTZER, AND M. TAN, *Sparse logistic regression with lp penalty for biomarker identification*, Statistical Applications in Genetics and Molecular Biology, 6 (2007). 1.3

[37] J. MARSDEN AND A. TROMBA, *Vector Calculus*, W.H. Freeman and Co., New York, New York, 1996. 3.7

[38] P. MCCULLAGH AND J. NELDER, *Generalized Linear Models*, Chapman and Hall, Boca Raton, Florida, 1989. 1.3, 2.2

[39] N. MEINSHAUSEN, *Relaxed lasso*, Computational Statistics & Data Analysis, 52 (2007), pp. 374 – 393. 3.6

[40] N. MEINSHAUSEN AND P. BÜHLMANN, *High-dimensional graphs and variable selection with the lasso*, Annals of Statistics, 34 (2006), pp. 1436–1462. 7.3

[41] A. B. OWEN, *A robust hybrid of lasso and ridge regression*, tech. report, Stanford University, 2006. 1.4, 7.5

[42] S. ROSSET AND J. ZHU, *Piecewise linear regularized solution paths*, Annals of Statistics, 35 (2007), pp. 1012–1030. 1.4

[43] F. SANTOSA AND W. W. SYMES, *Linear inversion of band-limited reflection seismograms*, SIAM Journal on Scientific Computing, 7 (1986), pp. 1307–1330. 1

[44] E. D. SCHIFANO, R. L. STRAWDERMAN, AND M. T. WELLS, *Majorization-minimization algorithms for nonsmoothly penalized objective functions*, Electronic Journal of Statistics, 4 (2010), pp. 1258–1299. 4, 4.2.1, 4.2.1.1, 7.5

[45] G. SCHWARZ, *Estimating the dimension of a model*, Annals of Statistics, 6 (1978), pp. 461–464. 3.6

[46] D. SCOTT, *Partial mixture estimation and outlier detection in data and regression*, in Theory and Applications of Recent Robust Methods, M. Hubert, G. Pison, A. Struyf, and S. V. Aelst, eds., Birkhauser, Basel, 2004, pp. 297–306. 2.1, 2.1.1, 7.5

[47] D. W. SCOTT, *Parametric statistical modeling by minimum integrated square error*, Technometrics, 43 (2001), pp. 274–285. 1.4, 2.1

[48] D. W. SCOTT, *Multivariate Density Estimation. Theory, Practice and Visualization*, John Wiley & Sons, Inc., 2008. 2.1

[49] D. W. SCOTT, *The L2E method*, Wiley Interdisciplinary Reviews: Computational Statistics, 1 (2009), pp. 45–51. 1.4, 2.1.1

[50] J. SHAO, *Mathematical Statistics*, Springer, second ed., July 2003. 1.1.1

[51] H. L. TAYLOR, S. C. BANKS, AND J. F. MCCOY, *Deconvolution with the l 1 norm*, Geophysics, 44 (1979), pp. 39–52. 1

[52] R. TIBSHIRANI, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological), 58 (1996), pp. pp. 267–288. 1, 1.1.3

[53] R. TIBSHIRANI, J. BIEN, J. FRIEDMAN, T. HASTIE, N. SIMON, J. TAYLOR, AND R. J. TIBSHIRANI, *Strong rules for discarding predictors in lasso-type problems.* arXiv:1011.2234v2, 2011. 7.6

[54] P. TSENG, *Convergence of a block coordinate descent method for nondifferentiable minimization*, Journal of Optimization Theory and Applications, 109 (2001), pp. 475–494. 10.1023/A:1017501703105. 4.3

[55] L. WANG, J. ZHU, AND H. ZOU, *Hybrid huberized support vector machines for microarray classification and gene selection*, Bioinformatics, 24 (2008), pp. 412–419. 1.4

[56] C. WOLF, K. MEISENHEIMER, M. KLEINHEINRICH, A. BORCH, S. DYE, M. GRAY, L. WISOTZKI, E. F. BELL, H. RIX, A. CIMATTI, G. HASINGER, AND G. SZOKOLY, *A catalogue of the Chandra Deep Field South with multi-colour classification and photometric redshifts from COMBO-17*, Astronomy and Astrophysics, 421 (2004), pp. 913–936. 6.1

[57] T. T. WU, Y. F. CHEN, T. HASTIE, E. SOBEL, AND K. LANGE, *Genomewide association analysis by lasso penalized logistic regression*, Bioinformatics, 25 (2009), pp. 714–721. 1.3, 7.6

[58] T. T. WU AND K. LANGE, *Coordinate descent algorithms for lasso penalized regression*, Annals of Applied Statistics, 2 (2008), pp. 224–244. 3

[59] F.-R. ZHANG, W. HUANG, S.-M. CHEN, L.-D. SUN, H. LIU, Y. LI, Y. CUI, X.-X. YAN, H.-T. YANG, R.-D. YANG, T.-S. CHU, C. ZHANG, L. ZHANG, J.-W. HAN, G.-Q. YU, C. QUAN, Y.-X. YU, Z. ZHANG, B.-Q. SHI, L.-H. ZHANG, H. CHENG, C.-Y. WANG, Y. LIN, H.-F. ZHENG, X.-A. FU, X.-B. ZUO, Q. WANG, H. LONG, Y.-P. SUN, Y.-L. CHENG, H.-Q. TIAN, F.-S. ZHOU, H.-X. LIU, W.-S. LU, S.-M. HE, W.-L. DU, M. SHEN, Q.-Y. JIN, Y. WANG, H.-Q. LOW, T. ERWIN, N.-H. YANG, J.-Y. LI, X. ZHAO, Y.-L. JIAO, L.-G. MAO, G. YIN, Z.-X. JIANG, X.-D. WANG, J.-P. YU, Z.-H. HU, C.-H. GONG, Y.-Q. LIU, R.-Y. LIU, D.-M. WANG, D. WEI, J.-X. LIU, W.-K. CAO, H.-Z. CAO, Y.-P. LI, W.-G. YAN, S.-Y. WEI, K.-J. WANG, M. L. HIBBERD, S. YANG, X.-J. ZHANG, AND J.-J. LIU, *Genomewide association study of leprosy*, New England Journal of Medicine, 361 (2009), pp. 2609–2618. 1

[60] P. ZHAO AND B. YU, *On model selection consistency of lasso*, J. Mach. Learn. Res., 7 (2006), pp. 2541–2563. 7.3

[61] H. ZHOU, D. ALEXANDER, AND K. LANGE, *A quasi-newton acceleration for high-dimensional optimization algorithms*, Statistics and Computing, (2009), pp. 1–13. 10.1007/s11222-009-9166-3. 7.6

[62] H. ZOU AND T. HASTIE, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 301–320. 1.1.3

[63] H. ZOU, T. HASTIE, AND R. TIBSHIRANI, *On the "degrees of freedom" of the lasso*, Annals of Statistics, 35 (2007), pp. 2173–2192. 3.6

[64] H. ZOU AND R. LI, *One-step sparse estimates in nonconcave penalized likelihood models*, Annals of Statistics, 36 (2008), pp. 1509–1533. 7.5