Robust Parametric Classification and Variable Selection by a Minimum Distance Criterion

Eric C. Chi* and David W. Scott†

Abstract

We investigate a robust penalized logistic regression algorithm based on a minimum distance criterion. Influential outliers are often associated with the explosion of parameter vector estimates, but in the context of standard logistic regression, the bias due to outliers always causes the parameter vector to implode, that is shrink towards the zero vector. Thus, using LASSO-like penalties to perform variable selection in the presence of outliers can result in missed detections of relevant covariates. We show that by choosing a minimum distance criterion together with an Elastic Net penalty, we can simultaneously find a parsimonious model and avoid estimation implosion even in the presence of many outliers in the important small n large p situation. Minimizing the penalized minimum distance criterion is a challenging problem due to its nonconvexity. To meet the challenge, we develop a simple and efficient MM algorithm that can be adapted gracefully to the small n large p context. Performance of our algorithm is evaluated on simulated and real data sets. This article has supplementary materials online.

Keywords: Logistic regression, Robust estimation, Implosion breakdown, LASSO, Elastic Net, Majorization-Minimization

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 have made it possible to comprehensively query a patient's transcriptional activity at a cellular level. 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.

*Eric C. Chi (E-mail: ecchi@ucla.edu) is Postdoctoral Scholar, Department of Human Genetics, University of California, Los Angeles CA 90095-7088.

†David W. Scott (E-mail: scottdw@rice.edu) is Professor, Department of Statistics, Rice University, Houston, TX 77005.

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" (Tibshirani, 1996; Chen et al., 1998). Indeed, ℓ_1-penalized maximum likelihood model fitting has inspired many related approaches that simultaneously do model fitting and variable selection. These approaches have been extended from linear regression to generalized linear models. In particular, linear models minimizing the logistic deviance loss with an Elastic Net penalty (Zou and Hastie, 2005) have been well studied (Genkin et al., 2007; Liu et al., 2007; Wu et al., 2009; Friedman et al., 2010)

Nonetheless while ℓ_1-penalized maximum likelihood methods have proved their worth at recovering parsimonious models, less attention has been given to extending these methods to handle outliers in high dimensional data. For example in biological data, tissue samples may be mislabeled or be contaminated. The majority of prior work centers on linear regression (Rosset and Zhu, 2007; Wang et al., 2007; Li et al., 2011; Alfons et al., 2012), although there are a few exceptions. Rosset and Zhu (2007) and Wang, Zhu, and Zou (2008) discuss using a Huberized hinge loss for regularized classification, and van de Geer (2008) studies LASSO penalization of generalized linear models. Nonetheless, with the exception of the ℓ_1-penalized least trimmed squares regression procedure of Alfons et al. (2012) and the Huberized hinge loss, these approaches can provide robustness only to outliers in the response variable, not to outliers in the covariates. Moreover, neither paper on the Huberized hinge loss is primarily concerned with robustness. Rosset and Zhu (2007) present impressive general conditions that ensure piecewise linear regularization paths. The Huberized hinge loss is introduced as an illustration and applied on a small example that highlights its prediction accuracy in the presence of a single gross outlier. Despite being introduced as a loss for a robust procedure in Rosset and Zhu (2007), the primary motivation for using the Huberized hinge loss in Wang et al. (2008) is the fast algorithm introduced in Rosset and Zhu (2007) for computing the entire regularization path, not its robustness properties. We will see later that this loss can struggle under a heavy dose of outliers.

Robustness against outlying covariate values warrants further investigation. It is not surprising that outliers may bias estimation. What is less well appreciated is that outliers can strongly influence variable selection. In this paper we identify some circumstances that motivate robust variants of penalized estimation and develop a minimum distance estimator for logistic regression.

To address the $n \ll p$ scenario when predictors are correlated we add the Elastic Net penalty. We evaluate the performance of our approach through simulated and real data.

Robust methods of logistic regression are not new in the classic $n > p$ case. 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 (Künsch et al., 1989; Carroll and Pederson, 1993) or on extreme predicted probabilities (Copas, 1988; Carroll and Pederson, 1993; Bianco and Yohai, 1996).

An alternative approach is to use minimum distance estimation (Donoho and Liu, 1988). The minimum distance estimator used in this paper can also be seen as a method that downweights the contributions of outliers (Chi, 2011). The work in Bondell (2005) is similar to ours in that he considered fitting parameters by minimizing a weighted Cramér-von Mises distance. The 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 outliers. Moreover, the robust loss function we propose has a particularly simple form which, when combined with the Elastic Net penalty, can be solved very efficiently for large problems by minimizing a series of penalized least squares problems with coordinate descent.

The rest of this paper is organized as follows. In Section 2 we review maximum likelihood estimation (MLE) of the logistic regression model and demonstrate the potentially deleterious effects of outliers on variable selection with the ℓ_1-penalized MLE. We introduce our robust loss function in Section 3. In Section 4 we describe algorithms for fitting our robust logistic regression model. In Sections 5 and 6 we present results on real and simulated data. Section 7 concludes with a summary of our work and also future directions.

2 Standard logistic regression and implosion breakdown

Throughout this paper we adopt the following conventions. We assume that the columns of the design matrix $\mathbf{X}$ are centered. We overload notation so that if f is a function of a scalar, then f evaluated at vector or matrix should be interpreted as being evaluated element-wise. For a linear model $\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}$ we will often employ the compact notations $\tilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X}) \in \mathbb{R}^{n\times(p+1)}$ and $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{p+1}$.

In binary regression, we seek to predict or explain an observed response $\mathbf{y} \in \{0,1\}^n$ using predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$, where $n \ll p$ may be expected. In typical expression microarray data we encounter $n \approx 100$ and $p \approx 10^4$, while with single nucleotide polymorphism (SNP) array data both n and p may be larger by a factor of 10. Let the conditional probabilities be given by $P(Y_i = 1|X_i = \mathbf{x}_i) = F(\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\theta})$ where $F(u) = 1/(1 + \exp(-u))$. Then under this assumption, in standard logistic regression (McCullagh and Nelder, 1989) we minimize the negative log-likelihood of a linear summary of the predictors,

$$\mathbf{y}^{\mathsf{T}}\tilde{\mathbf{X}}\boldsymbol{\theta} - \mathbf{1}^{\mathsf{T}}\log(\mathbf{1} + \exp(\tilde{\mathbf{X}}\boldsymbol{\theta})). \tag{2.1}$$

A simple univariate example illustrates the bias that outliers can introduce into this estimation procedure. In the top panel of Figure 1 we see that the addition of 5 and 10 outliers among the controls shrinks $\hat{\boldsymbol{\beta}}$ towards zero. In fact, Croux et al. (2002) showed that with p covariates only $2p$ such outliers are required to make $\|\hat{\boldsymbol{\beta}}\|_2 < \epsilon$ for any desired ϵ. Our robust estimator, which we introduce in the next section, produces virtually the same curves shown in the bottom panel of Figure 1.

This "implosion" breakdown phenomenon has implications for LASSO based variable selection. Consider what happens when we add 999 noise covariates which are independent of the class labels to the scenario depicted in Figure 2 and perform ℓ_1-penalized logistic regression. The top panel of Figure 2 shows the corresponding regularization paths or the values of the fitted regression coefficients as a function of the penalization parameter. As outliers are added the regularization path for the relevant covariate X_1 quickly falls into the noise.

The LASSO performs continuous variable selection by shrinking to zero regression coefficients of covariates with very low correlation with the responses. If outliers are present in relevant covariates, then the combination of implosion breakdown and soft-thresholding by the LASSO can lead to missed detection of relevant covariates. In contrast we see in the bottom panel of Figure 2 that the corresponding regularization paths obtained using our robust estimator are insensitive to outliers and so relevant covariates still have the chance of being selected. This simple example highlights the potential importance of penalized robust estimation procedures. In the next section we describe our robust estimator.

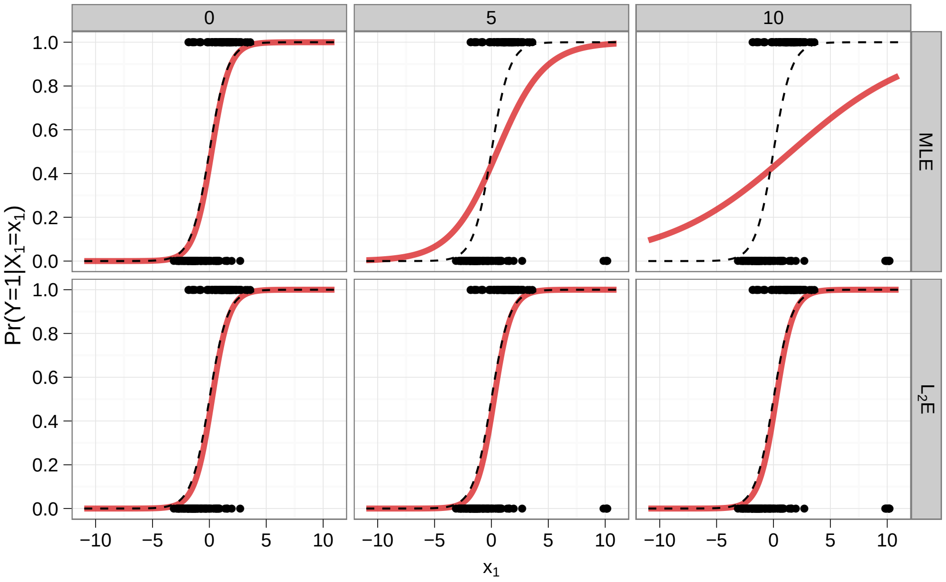

Figure 1: Univariate regression onto X_1. The dashed line denotes the logistic model that generated the data; the heavy solid line denotes the estimated response. The number of outliers (0, 5, 10) increases from left to right. The first row shows MLE results; the second shows L$_2$E results.

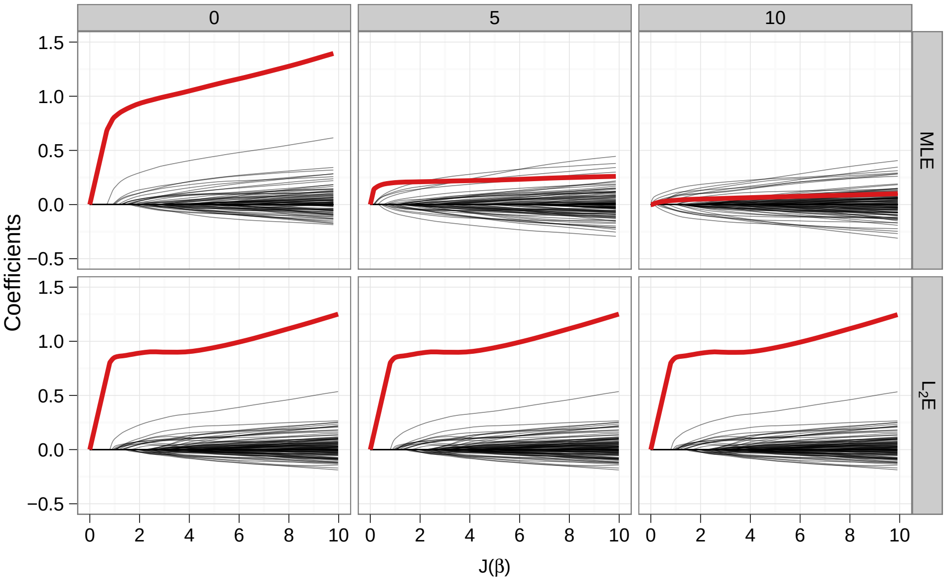

Figure 2: Regularization paths. The heavy line denotes the path for the relevant regression coefficient β_1; $J(\boldsymbol{\beta})$ is the 1-norm of $\boldsymbol{\beta}$. The number of outliers (0, 5, 10) increases from left to right; 999 irrelevant covariates have been added. The first row shows MLE results; the second shows L$_2$E results.

3 The Minimum Distance Estimator

Let $P_{\boldsymbol{\theta}}$ be a probability mass function (PMF), specified by a parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$, believed to be generating data $Y_1, \ldots, Y_n$ that take on values in the discrete set χ. Let P be the unknown true PMF generating the data. 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 L$_2$ distance between $P_{\boldsymbol{\theta}}$ and P. Thus, we pose the following variational optimization problem; we seek $\hat{\boldsymbol{\theta}} \in \Theta$ that minimizes

$$\sum_{y \in \chi} \left[P_{\boldsymbol{\theta}}(y) - P(y)\right]^2. \tag{3.1}$$

Although finding such a $\boldsymbol{\theta}$ is impossible since P is unknown, it is possible to find a $\boldsymbol{\theta}$ that minimizes an unbiased estimate of this distance. Expanding the sum in (3.1) gives us

$$\sum_{y \in \chi} P_{\boldsymbol{\theta}}(y)^2 - 2\sum_{y \in \chi} P_{\boldsymbol{\theta}}(y)P(y) + \sum_{y \in \chi} P(y)^2.$$

The second summation is an expectation $E[P_{\boldsymbol{\theta}}(Y)]$ where Y is a random variable drawn from P. This summation can be estimated from the data by the sample mean. The third summation does not depend on $\boldsymbol{\theta}$. With these observations in mind, we use the following fully data-based loss function

$$L(\boldsymbol{\theta}) = \sum_{y \in \chi} P_{\boldsymbol{\theta}}(y)^2 - \frac{2}{n}\sum_{i=1}^{n} P_{\boldsymbol{\theta}}(y_i) \tag{3.2}$$

and seek a $\hat{\boldsymbol{\theta}}$ such that $L(\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$. The estimate $\hat{\boldsymbol{\theta}}$ is called an L$_2$ estimate or L$_2$E in Scott (2001).

The above minimization problem is a familiar one associated with bandwidth selection for histograms and more generally for kernel density estimators (Scott, 1992). Applying a commonly used criterion in nonparametric 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. (1998) introduced a family of divergences which includes the L$_2$E 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. The MLE is the most efficient but least robust member in this family of estimation procedures. The L$_2$E represents a reasonable tradeoff between efficiency and robustness. Scott (2001, 2004) demonstrated that the L$_2$E has two benefits, the aforementioned robustness properties and computational tractability. The tradeoff in asymptotic efficiency is similar to that seen in comparing the mean and median as a location estimator. Indeed, while other members in this family may possess a better tradeoff, the L_2E has the advantage of admitting a simple and fast computational solution as we will show in Section 4.

We now show that the L_2E method applied to logistic regression amounts to solving a non-linear least squares problem. We seek to minimize a surrogate measure of the L_2 distance between the logistic conditional probability and the conditional probability generating the data. If the $\mathbf{x}_i$ are unique, then $y_i \sim \mathrm{B}(1, p_i)$ where $p_i = F(\tilde{\mathbf{x}}_i^\mathsf{T}\boldsymbol{\theta})$. The L_2E loss for this one sample is $p_i^2 + (1-p_i)^2 - 2[y_i p_i + (1-y_i)(1-p_i)]$. Extending to the entire sample, a sensible approach is to minimize the average L_2 distance, namely

$$\frac{1}{n}\sum_{i=1}^{n}\left[p_i^2 + (1-p_i)^2 - 2[y_i p_i + (1-y_i)(1-p_i)]\right]. \tag{3.3}$$

Up to an additive constant that does not depend on $\boldsymbol{\theta}$, the criterion in (3.3) can be compactly written as

$$L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) = \frac{1}{n}\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2,$$

after dividing by two. Remarkably, minimizing this unassuming loss function produces robust logistic regression coefficients. A closer inspection of the estimating equations gives some intuition for the logistic L_2E's robustness. A stationary point $\boldsymbol{\theta}^*$ of the L_2E loss satisfies

$$0 = \sum_{i=1}^{n}\gamma_i^*\mathbf{x}_i[y_i - F(\tilde{\mathbf{x}}_i^\mathsf{T}\boldsymbol{\theta}^*)]$$

where $\gamma_i^* = F(\tilde{\mathbf{x}}_i^\mathsf{T}\boldsymbol{\theta}^*)[1 - F(\tilde{\mathbf{x}}_i^\mathsf{T}\boldsymbol{\theta}^*)]$. Thus, at a stationary point $\boldsymbol{\theta}^*$, the discrepancies between observed and fitted values, namely $y_i - F(\tilde{\mathbf{x}}_i^\mathsf{T}\boldsymbol{\theta}^*)$, are small for samples with predicted values that are far from the extreme values of one and zero, namely samples for which γ_i^* are not close to zero. The ith discrepancy is free to be large for samples with predicted values close to zero or one, namely samples for which γ_i^* are close to zero. Very large and small predicted values tend to occur at extreme values of the covariates given the sigmoid shape of F. Thus, observations that are extreme in the covariate space contribute very little to the estimating equations at $\boldsymbol{\theta}^*$. Moreover, we see that the robustness does not rely on F being the logistic link; rather we just require that F be sigmoid. Finally, we note that the estimating equations also show us that the L_2E is affine equivariant, namely linear transformations of the covariates change the estimated regression coefficients accordingly, and therefore linear transformations of the covariates do not change the fitted responses. For more in depth discussion on the theory behind minimum distance estimators like the L$_2$E, we refer readers to the works of Basu et al. (1998) and Donoho and Liu (1988).

Before moving on to discuss our algorithm, we remark that the L$_2$ distance has been used before for classification problems. Kim and Scott (2008, 2010) used the L$_2$ distance to perform classification using kernel density estimates. Their application of the L$_2$ distance, however, is more in line with its customary use in nonparametric density estimation whereas we use it to robustly fit a parametric model.

4 Estimation with convex quadratic majorizations

We now derive an algorithm for finding the logistic L$_2$E solution by minimizing a series of convex quadratic losses. We minimize the L$_2$E loss with a Majorization-Minimization (MM) algorithm (Lange, Hunter, and Yang, 2000; Hunter and Lange, 2004) because it is numerically stable and easy to implement. Most importantly, our MM algorithm is also easily adapted to handle LASSO-like penalties.

The strategy behind MM algorithms 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. Formally stated, a real-valued function h majorizes a real-valued function g at $\mathbf{v}$ if $h(\mathbf{u}) \geq g(\mathbf{u})$ for all $\mathbf{u}$ and $h(\mathbf{v}) = g(\mathbf{v})$. Given a procedure for constructing a majorization, we can define the MM algorithm to find a minimizer of a function g as follows. Let $\mathbf{v}^{(k)}$ denote the kth iterate: (1) find a majorization $h(\mathbf{v}; \mathbf{v}^{(k)})$ of g at $\mathbf{v}^{(k)}$; (2) set $\mathbf{v}^{(k+1)} = \arg\min_{\mathbf{v}} h(\mathbf{v}; \mathbf{v}^{(k)})$; and (3) repeat until convergence. This algorithm always takes non-increasing steps with respect to g. By using the MM algorithm, we can convert a hard optimization problem into a series of simpler ones, each of which is easier to minimize than the original.

To estimate $\hat{\boldsymbol{\theta}}$ such that $L(\mathbf{y}, \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}) = \min_{\boldsymbol{\theta}} L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ we rely on the following convex quadratic majorization.

Theorem 4.1. *The following function majorizes $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$:*

$$L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) = L(\mathbf{y}, \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) + \frac{2}{n}\mathbf{z}_{\tilde{\boldsymbol{\theta}}}^{\mathsf{T}}\tilde{\mathbf{X}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \frac{\eta}{n}\|\tilde{\mathbf{X}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\|_2^2, \tag{4.1}$$

where $\mathbf{z}_{\tilde{\boldsymbol{\theta}}} = 2\mathbf{G}[F(\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) - \mathbf{y}]$, $\mathbf{G}$ is diagonal with $g_{ii} = F(\tilde{\mathbf{x}}_i^{\mathsf{T}}\tilde{\boldsymbol{\theta}})[1 - F(\tilde{\mathbf{x}}_i^{\mathsf{T}}\tilde{\boldsymbol{\theta}})]$, and $\eta > 0$ is sufficiently large.

Using the majorization (4.1) in an MM algorithm results in iterative least squares. A proof of Theorem 4.1 is given in the Supplementary Materials. We are able to find a simple convex quadratic majorization since the logistic L$_2$E loss has bounded curvature. A sharp lower bound on η is given by the maximum curvature of the logistic L$_2$E loss over all parameter values. The bound is derived in the Supplementary Materials. The practical implication is 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.

We can express the majorization $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ in (4.1) as

$$L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) = \eta(\tilde{\beta}_0 - \beta_0 - \frac{1}{\eta}\bar{z}_{\tilde{\boldsymbol{\theta}}})^2 + \frac{\eta}{n}\|\zeta(\tilde{\boldsymbol{\theta}}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + K(\tilde{\boldsymbol{\theta}}),$$

where $\bar{z}_{\tilde{\boldsymbol{\theta}}} = n^{-1}\mathbf{1}^{\mathsf{T}}\mathbf{z}_{\tilde{\boldsymbol{\theta}}}$, $\zeta(\tilde{\boldsymbol{\theta}}) = \mathbf{X}\tilde{\boldsymbol{\beta}} - \eta^{-1}(z_{\tilde{\boldsymbol{\theta}}} - \bar{z}_{\tilde{\boldsymbol{\theta}}}\mathbf{1})$, and $K(\tilde{\boldsymbol{\theta}})$ is a constant that does not depend on $\boldsymbol{\theta}$. When $\mathbf{X}$ is full rank, as is often the case when $n > p$, then the solution to the normal equations is unique and the parameter updates are given by

$$\begin{aligned} \beta_0^{(m+1)} &= \beta_0^{(m)} - \eta^{-1}\bar{z}_{\boldsymbol{\theta}^{(m)}}, \\ \boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} - \frac{1}{\eta}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{z}_{\boldsymbol{\theta}^{(m)}}. \end{aligned} \tag{4.2}$$

The descent direction has a simple update since the Hessian approximation is computed only once for all iterations.

The majorization given in Theorem 4.1 can be adapted for regularization. It follows immediately that $(1/2)L(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}}) + \lambda J(\boldsymbol{\beta})$ majorizes $(1/2)L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) + \lambda J(\boldsymbol{\beta})$ for a penalty function $J : \mathbb{R}^p \to \mathbb{R}_+$ and positive regularization parameter λ. Note that the intercept parameter is not penalized. Regularization is useful for stabilizing estimation procedures. For example, if $\mathbf{X}$ is not full rank or has a large condition number, a ridge penalty can salvage the situation. We then seek the minimizer to the following problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{p+1}} \frac{1}{2n}\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2 + \lambda\frac{1}{2}\|\boldsymbol{\beta}\|_2^2,$$

which we can solve by minimizing the majorization $L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) + \lambda \|\boldsymbol{\beta}\|_2^2$. Since the intercept is not penalized, the intercept updates are the same as in (4.2). The update for $\boldsymbol{\beta}$ becomes

$$\boldsymbol{\beta}^{(m+1)} = \boldsymbol{\beta}^{(m)} - \frac{1}{\eta}(\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\mathbf{z}_{\boldsymbol{\theta}^{(m)}}. \tag{4.3}$$

Under suitable regularity conditions, the MM algorithm for solving the ridge penalized logistic L$_2$E problem is guaranteed to converge to a stationary point of $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta}) + \lambda\|\boldsymbol{\beta}\|_2^2$. This follows from global convergence properties of MM algorithms that involve continuously differentiable objective and majorization functions (Lange, 2010). On the other hand, the MM algorithm for the unregularized version of the problem is not guaranteed to converge based on the sufficient conditions given in Lange (2010) because the objective function is not coercive (i.e., not all its level sets are compact) and the quadratic majorization is not strictly convex in $\boldsymbol{\theta}$ unless $\mathbf{X}$ is full rank. Adding the ridge penalty remedies both situations, and sufficient conditions for global convergence are met.

Another reason to consider regularization is to perform continuous variable selection via a LASSO-like penalty. In particular, consider the penalized majorizer for the L$_2$E loss regularized by the Elastic Net penalty, $J(\boldsymbol{\beta}) = \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)/2\|\boldsymbol{\beta}\|_2^2\right)$ where $\alpha \in [0,1]$ is a mixing parameter between the ridge and LASSO penalty. Since our work is motivated by genomic data which are 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 (Zou and Hastie, 2005). The LASSO, in contrast, tends to select one covariate among a group correlated covariates and exclude the rest. If groupings among the covariates are known in advance, a group LASSO penalty could be used (Yuan and Lin, 2006). The Elastic Net penalty is useful in that it performs group selection without prespecification of the groups. Thus, we are interested in generating MM iterates $\boldsymbol{\theta}^{(m)} = \left(\beta_0^{(m)}, \boldsymbol{\beta}^{(m)}\right)$ where

$$\begin{aligned}
\beta_0^{(m+1)} &= \beta_0^{(m)} - \eta^{-1}\overline{z}_{\boldsymbol{\theta}^{(m)}} \\
\boldsymbol{\beta}^{(m+1)} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ \frac{\eta}{2n}\|\zeta(\boldsymbol{\theta}^{(m)}) - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right).
\end{aligned} \tag{4.4}$$

Before discussing how to practically solve the surrogate minimization problem, note that regardless of how (4.4) is solved, we have the following guarantee on the convergence of the MM iterates.

Theorem 4.2. *Under suitable regularity conditions, for any starting point $\boldsymbol{\theta}^{(0)}$, the sequence of iterates $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ generated by (4.4) converges to a stationary point of*

$$\frac{1}{2n}\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{(1-\alpha)}{2}\|\boldsymbol{\beta}\|_2^2\right),$$

where $\lambda > 0$ and $\alpha \in [0, 1)$.

A proof is given in the Supplementary Materials and relies on an extension of the global convergence properties of MM algorithms for locally Lipschitz continuous objective and majorization functions (Schifano et al., 2010). Note that Theorem 4.2 restricts $\alpha < 1$, i.e., algorithmic convergence of the LASSO regularized logistic L_2E is not guaranteed. This condition is imposed to ensure that the majorization is strictly convex in $\boldsymbol{\beta}$. In our experience, the LASSO regularized logistic L_2E does not have algorithmic convergence issues in practice. As a final remark on algorithmic convergence, note that since the ridge penalty is a special case of the Elastic Net, Theorem 4.2 implies that ridge penalized logistic L_2E (4.3) will also converge.

To solve (4.4) we turn to coordinate descent which has been shown to efficiently solve penalized regression problems when selecting relatively few groups of correlated predictors (Friedman, Hastie, Höfling, and Tibshirani, 2007; Wu and Lange, 2008). 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.

The jth coordinate update during the kth round of coordinate descent of the mth MM iteration, denoted $\beta_j^{(m,k)}$, has a simple form (Donoho and Johnstone, 1995) and is given by the subgradient equations to be

$$\beta_j^{(m,k)} = \frac{S\left(\frac{\eta}{n}\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{r}^{(m,k,j)}, \lambda\alpha\right)}{\frac{\eta}{n}\|\mathbf{x}_{(j)}\|_2^2 + \lambda(1-\alpha)},$$

where $\mathbf{x}_{(j)}$ denotes the jth column of $\mathbf{X}$ and $\mathbf{r}^{(m,k,j)}$ is a vector of partial residuals with ith entry

$$r_i^{(m,k,j)} = \zeta_i(\boldsymbol{\theta}^{(m)}) - \left(\sum_{j'=1}^{j-1} x_{ij'}\beta_{j'}^{(m,k)} + \sum_{j'=j+1}^{p} x_{ij'}\beta_{j'}^{(m,k-1)}\right),$$

and S is the soft-threshold function: $S(a, \lambda) = \operatorname{sign}(a)\max(|a| - \lambda, 0)$. Additional details on how coordinate descent is nested within the MM steps and how convergence is evaluated can be found in the Supplementary Materials.

5 Simulations

In this section we report on three simulations comparing the MLE and $\mathrm{L}_2\mathrm{E}$ results. 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 by the likelihood in (2.1) with parameters $\beta_0 = 0$ and $\boldsymbol{\beta} = (1, 0.5, 1, 2)^\mathsf{T}$. The covariates $\mathbf{x}_i$ were drawn from one of two populations. For $i = 1, \ldots, 100$, the $\mathbf{x}_i$ are i.i.d samples from $N(\boldsymbol{\mu}, 0.16\,\mathbf{I}_p)$ and for $i = 101, \ldots 200$, they are i.i.d samples from $N(-\boldsymbol{\mu}, 0.16\,\mathbf{I}_p)$, where $p = 4$ and $\boldsymbol{\mu} = (0.25, 0.25, 0.25, 0.25)^\mathsf{T}$. The responses were generated independently as $y_i \sim \mathrm{B}(1, F(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}))$.

5.1 Estimation in Low Dimensions

In the first scenario, we added a single outlier, $(y_{201}, \mathbf{x}_{201})$ where $y_{201} = 0$ and $\mathbf{x}_{201} = (\delta, \delta, \delta, \delta)^\mathsf{T}$ and δ took on values in $\{-0.25, 1.5, 3, 6, 12, 24\}$. In words, the 201st point was moved in covariate space along the line that runs through the centroids of the two subpopulations. In the second scenario, we added a variable number of outliers at a single location: $\{(y_i, \mathbf{x}_i)\}_{i=201}^N$, where $y_i = 0$ and $\mathbf{x}_i = (3, 3, 3, 3)^\mathsf{T}$ for $i = 201, \ldots, N$ and the number of outliers is $N = 0, 1, 5, 10, 15, 20$. For each sequence of scenarios described, we performed logistic regression and $\mathrm{L}_2\mathrm{E}$ regression. Figures 3 and 4 summarize the results of first and second scenario, respectively.

The results show two features of the $\mathrm{L}_2\mathrm{E}$ versus the MLE. Consider the first scenario. Figure 3 shows how $\|\hat{\boldsymbol{\beta}}\|_2$ under each estimation procedure varies with the position of outlier is moved. The MLE values suffer from implosion breakdown as the 201st point is moved from -0.25 to 24, i.e., $\|\hat{\boldsymbol{\beta}}\|_2$ tends towards 0 as the leverage of the 201st point increases. In contrast, the $\mathrm{L}_2\mathrm{E}$ is insensitive to the placement of the 201st point. The second observation is that the $\mathrm{L}_2\mathrm{E}$'s unbiasedness comes at the cost of increased variance. The $\mathrm{L}_2\mathrm{E}$'s spread is greater than the MLE's for all locations of the outlier. Similar behavior is observed in the second scenario. Figure 4 shows that implosion breakdown ensues as outliers are added at fixed position. Detailed numerical summaries of the fitted coefficients (sample mean, standard deviation, estimated mean squared error) of these experiments can be found in the Supplementary Materials.

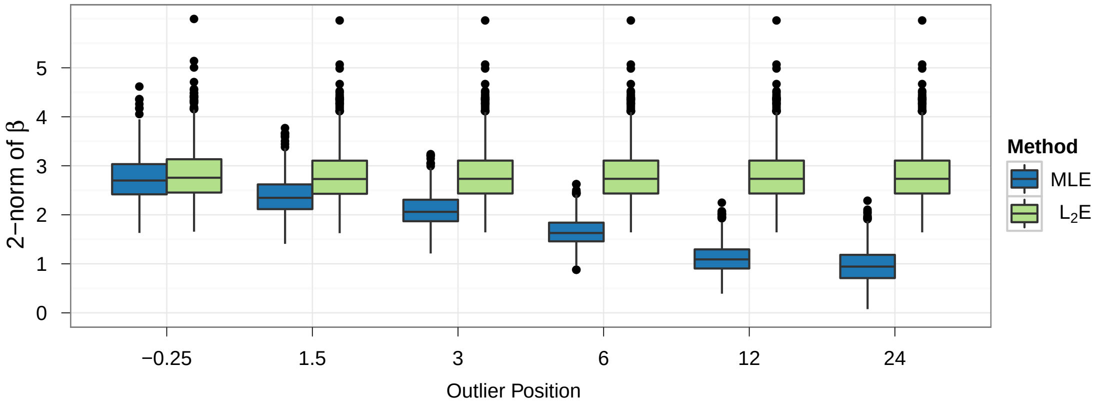

Figure 3: The 2-norm of the regression coefficients (intercept not included) as a function of the position of the single outlier.

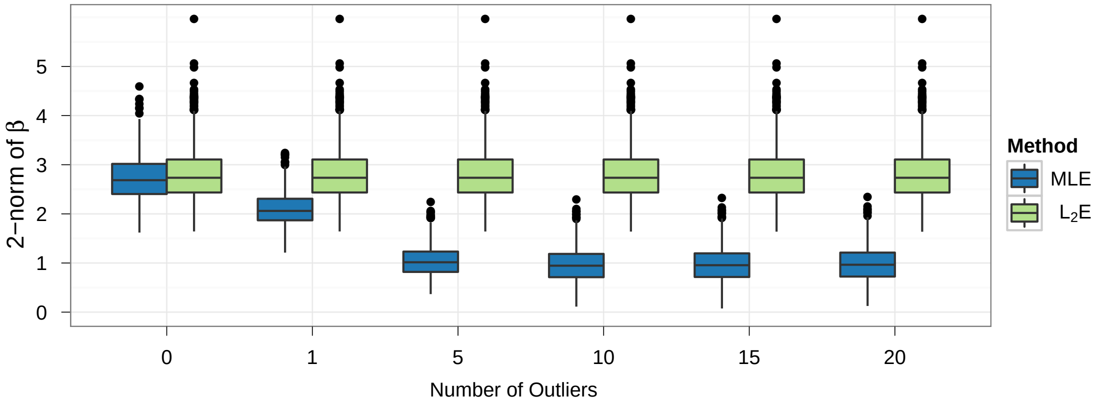

Figure 4: The 2-norm of the regression coefficients (intercept not included) as a function of the number of outliers at a fixed position.

5.2 Variable Selection in High Dimensions

In the variable selection experiment we considered a high dimensional variation on the first scenario. We generated 10 data sets each with $n = 500$ observations. The covariates were drawn from one of three multivariate normal populations. For $i = 1, \dots 200$, the $\mathbf{x}_i$ are i.i.d. samples from $N(\boldsymbol{\mu}, 0.75\,\mathbf{I}_p)$. For $i = 201, \dots, 400$, the $\mathbf{x}_i$ are i.i.d. samples from $N(-\boldsymbol{\mu}, 0.75\,\mathbf{I}_p)$. For $i = 401, \dots, 500$, the $\mathbf{x}_i$ are i.i.d. samples from $N(\boldsymbol{\nu}, 0.25\,\mathbf{I}_p)$ where $p = 500$, $\mu_i = 0.3$ for $i = 1, \dots, 50$ and $\mu_i = 0$ for $i = 51, \dots, 500$, and $\nu_i = 1$ for $i = 1, \dots, 50$ and $\nu_i = 0$ for $i = 51, \dots, 500$. For $i = 1, \dots, 400$, the responses were generated independently as $y_i \sim \mathrm{B}(1, F(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}))$, where $\beta_0 = 0$ and $\boldsymbol{\beta} \in \mathbb{R}^{500}$ with $\beta_i = 1$ for $i = 1, \dots 50$ and $\beta_i = 0$ for $i = 51, \dots, 500$. For $i = 401, \dots, 500$, the responses were set to $y_i = 0$,

We then performed Elastic Net penalized regression ($\alpha = 0.6$) with the MLE and L$_2$E. Before continuing we note that there are two practical issues that need to be addressed, namely how to choose initial starting points since the optimization problem is not convex and how to choose the amount of penalization. In the Supplementary Materials, we describe in detail a heuristic for choosing the initial starting point based on the Karush-Kuhn-Tucker conditions of the optimization problem as well as a robust cross validation scheme for choosing the regularization parameter λ. To perform the Elastic Net penalized logistic regression we used the **glmnet** package in R (Friedman et al., 2010). We also compared the robust classifier of Wang et al. (2008) - the Hybrid Huberized Support Vector Machine (HHSVM) using an MM algorithm. Wang et al. (2008) provide details of the implementation and code for computing the solution paths of the HHSVM. However, their algorithm calculates the paths for a varying LASSO regularization parameter with a fixed ridge regularization parameter because they can be computed quickly by exploiting the piece-wise linearity of the paths under that parameterization of the Elastic Net. Our HHSVM implementation calculates regularization paths using the Elastic Net parameterization used in this article. Details on our implementation can be found in the Supplementary Materials.

Tables 1 and 2 show the number of true positives and false positives respectively for each method. We see that in scenarios of heavy contamination the L$_2$E demonstrates superior sensitivity and specificity compared to both the MLE and HHSVM. It is interesting to note that the MLE tends to be more sensitive than the HHSVM, but at a cost of being drastically less specific. For a closer look comparing the three methods, the cross-validation curves and regularization paths for

Table 1: True positive count with $n = p = 500$ and 50 nonzero covariates. L_2E is the most sensitive method. HHSVM is the least sensitive method.

	Replicate									
	1	2	3	4	5	6	7	8	9	10
MLE	14	10	8	10	1	10	0	14	11	15
HHSVM	1	3	2	2	1	2	1	2	4	2
L_2E	48	47	48	49	48	48	49	46	48	49

Table 2: False positive count with $n = p = 500$ and 50 nonzero covariates. L_2E is the most specific method. MLE is the least specific method.

	Replicate									
	1	2	3	4	5	6	7	8	9	10
MLE	141	95	56	148	0	141	0	128	136	170
HHSVM	0	4	1	1	1	0	1	0	0	0
L_2E	0	0	2	0	0	0	1	1	0	1

a replicate can be found in the Supplementary Materials.

6 Real data examples

6.1 An $n > p$ example: Predicting abnormal and normal vertebral columns

We first consider a real data set in the $n > p$ regime. We present results on the vertebral column data set from the UCI machine learning repository, as described by Frank and Asuncion (2010). The data set consists of 310 patients which have been classified as belonging to one of three groups: Normal (100 patients), Disk Hernia (60 patients), Spondylolisthesis (150 patients). In addition to a classification label, six predictor variables are recorded for each patient: pelvic incidence (PI), pelvic tilt (PT), lumbar lordosis angle (LLA), sacral slope (SS), pelvic radius (PR) and grade of spondylolisthesis (GS). All six predictor variables are continuous valued.

Table 3: Correlations among the six biomechanical attributes in the vertebrae data set.

	PI	PT	LLA	SS	PR	GS
PI	1.00	0.63	0.72	0.81	-0.25	0.64
PT	–	1.00	0.43	0.06	0.03	0.40
LLA	–	–	1.00	0.60	-0.08	0.53
SS	–	–	–	1.00	-0.34	0.52
PR	–	–	–	–	1.00	-0.03
GS	–	–	–	–	–	1.00

We consider the two class problem of discriminating normal vertebral columns from abnormal ones (Disk Hernia and Spondylolisthesis). Figure 5 plots the values of individual covariates for each patient. Table 3 shows the correlations between pairs of attributes. Note that the attributes for Disk Hernia and Normal patients overlap a good deal. We may expect similar results as seen in the second simulation scenario described in Section 5.1 where Disk Hernia patients play the role of a cluster of outlying observations. Due to the correlation, however, the outlying observations are not as distinctly outlying as seen in the simulation examples of Section 5.1. Consequently, it also might be anticipated that there will not be differences between the MLE and L_2E regularization paths. Indeed, Figure 6 shows the resulting regularization paths generated by the MLE and logistic L_2E for $\alpha = 0.2$. The paths are very similar for both methods for other values of α and are not shown. Different initial starting points did not change the resulting logistic L_2E regularization paths.

6.2 An $n \ll p$ example: A genome wide association study

We examine the lung cancer data of Amos et al. (2008). 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.

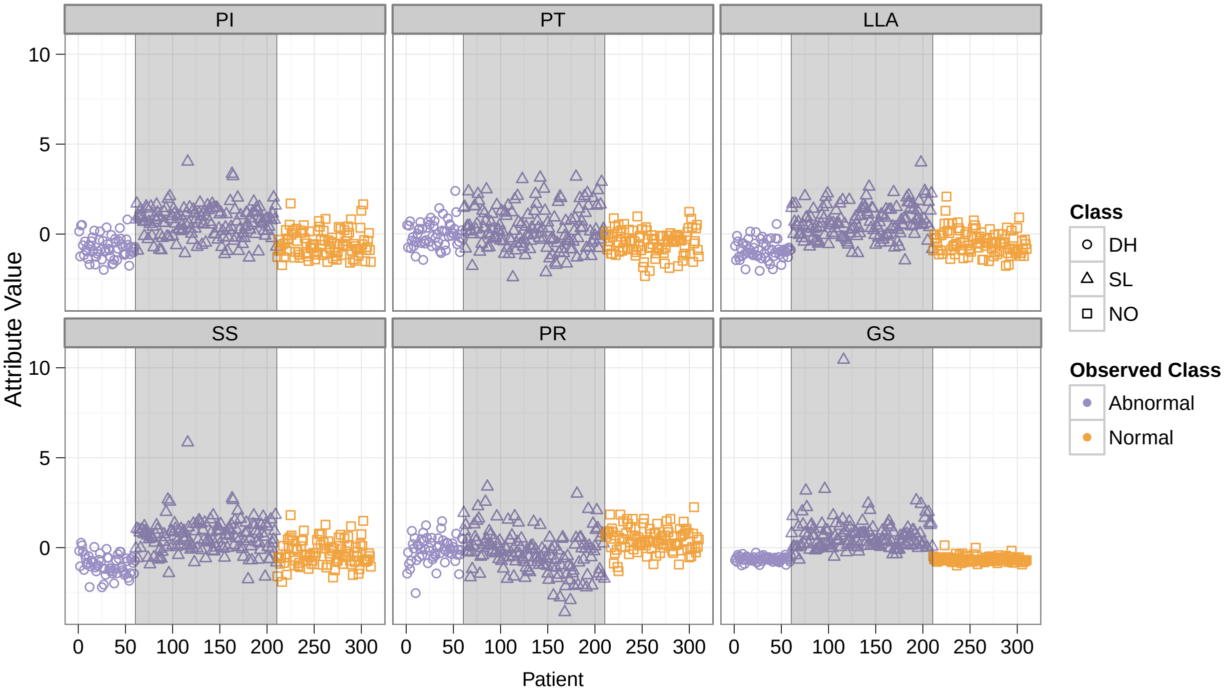

Figure 5: Dot plots of biomechanical attribute values for patients belonging to one of three classes. Patients are randomly ordered within their classes. DH and SL are lumped into the observed class Abnormal. Patients with SL (61 to 210) occupy the plot within the lightly shaded band.

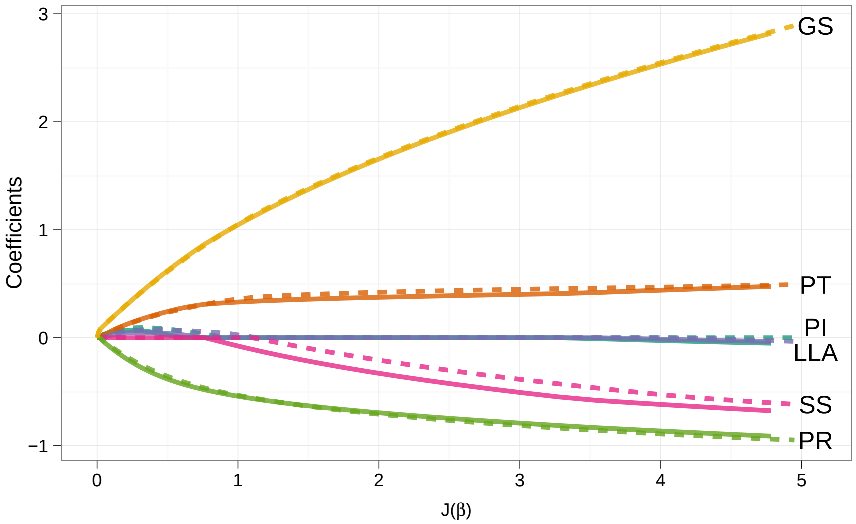

Figure 6: The regularization ($\alpha = 0.2$) paths for the MLE (solid) and L_2E (dashed) are not identical but very similar for the six biomechanical attributes in the vertebrae data set.

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. 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 disequilibrium. The pair rs1051730 and rs8034191 for example are in "high" linkage disequilibrium.

In this section we reexamine the discovery data using logistic $\mathrm{L}_2\mathrm{E}$ and the logistic MLE. Note that it is current practice of geneticists to do univariate inference with an adjustment for multiple testing and this approach was taken in Amos et al. (2008). Taking a multivariate approach as will be done in this section, however, allows the analyst to take into account dependencies between the SNPs. As an initial comparison we consider a subset of the entire data set and restrict our analysis to SNPs on chromosome 15. We impute missing genotypes at a SNP by using the MACH 1.0 package, a Markov Chain based haplotyper (Li, Ding, and Abecasis, 2006). After missing data are 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.

Figure 7 summarizes the variable selection results for the logistic $\mathrm{L}_2\mathrm{E}$ and MLE for $\alpha = 0.05, 0.5,$ and 0.95. There are three things to note. First, the regularization paths for the $\mathrm{L}_2\mathrm{E}$ and MLE are almost identical. Second, both methods produce regularization paths that identify rs1051730 (light-thick line) and rs8034191 (dark-thick line) as having the greatest partial correlation with the case/control status. Third, 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.

7 Discussion

Outliers can introduce bias in some commonly used maximum likelihood estimation procedures. This well known fact, however, warrants attention because bias can have material effects on the ubiquitous LASSO-based variable selection procedures. In the context of standard logistic regression, influential outliers cause implosion breakdown. In this paper we have demonstrated that the combination of implosion breakdown and the soft-thresholding mechanism of LASSO

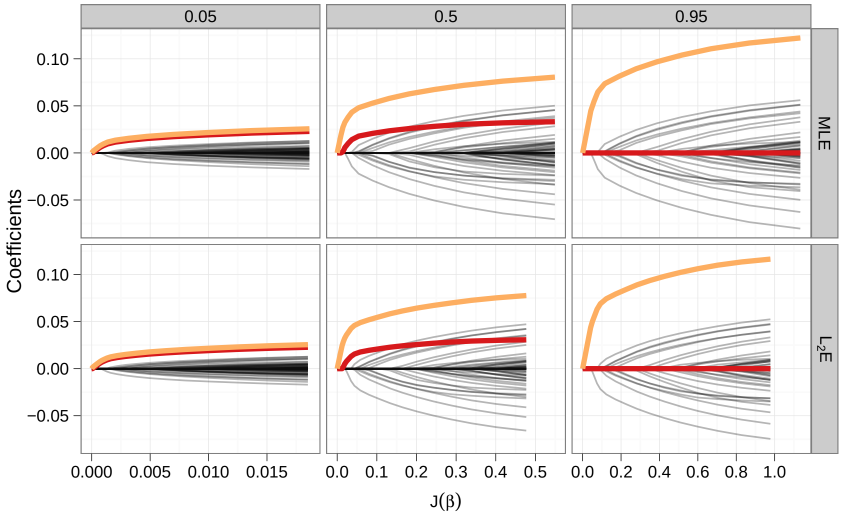

Figure 7: Regularization paths of regression coefficients of SNP markers on Chromosome 15 for L_2E and MLE for $\alpha = 0.05, 0.5$, and 0.95. The regularization paths for rs1051730 are in light-thick lines; the paths for rs8034191 are in dark-thick lines. The L_2E and MLE paths are nearly identical. For $\alpha = 0.95$, i.e. nearly LASSO regression, rs8034191 was not selected for the shown range of penalizations by either method.

variable selection can lead to missed detection of relevant predictors.

To guard against the undue influence of outliers on estimation and variable selection for binary responses, we propose a robust method for performing sparse logistic regression. Our method is based on minimizing the estimated L_2 distance between the logistic parametric model and the underlying true conditional distribution. The resulting optimization problem is a penalized non-linear least squares problem which we solve with an MM algorithm. Our MM algorithm in turn reduces the optimization problem to solving a series penalized least squares problems whose solution paths can be solved very efficiently with coordinate descent and warm starts.

Although we present our work as a method for robust binary logistic regression, our method immediately extends to other related contexts. Our algorithm can be extended to handle more than two classes. The generalization to the K-class multinomial is straightforward.

$$L(\mathbf{Y}, \tilde{\mathbf{X}}\boldsymbol{\Theta}) = \sum_{k=1}^{K} \lVert \mathbf{y}_k - F_k(\tilde{\mathbf{X}}\boldsymbol{\Theta}) \rVert_2^2,$$

where $y_{ik} = 1$ if the ith observation belongs to class k and 0 otherwise and the ith element of vector $F_k(\tilde{\mathbf{X}}\boldsymbol{\Theta})$ is given by

$$\frac{\exp(\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\theta}_k)}{1 + \sum_{j=1}^{K} \exp(\tilde{\mathbf{x}}_i^{\mathsf{T}}\boldsymbol{\theta}_j)}.$$

This non-linear least squares problem also has bounded curvature and consequently can also be solved by minimizing a sequence of LASSO-penalized least squares problems.

Our algorithm can also be used as a subroutine in performing robust binary principal component analysis and, more generally, robust binary tensor decompositions. A common strategy in array decompositions for multiway data, including multiway binary data, is to use block coordinate descent or alternating minimization (Collins, Dasgupta, and Schapire, 2001; Kolda and Bader, 2009; Lee, Huang, and Hu, 2010). For binary multiway data, each block minimization would perform a batch of independent robust logistic regressions.

We want to make clear that the logistic $\mathrm{L_2E}$ is not a competitor to the MLE but rather a complement. Both methods are computationally feasible and can be run on data together. As seen in the real data examples of Section 6, sometimes the logistic $\mathrm{L_2E}$ recovers the MLE solution. On the other hand, when discrepancies do occur, taking the MLE and $\mathrm{L_2E}$ solutions together can provide insight into the data that would be harder to identify with the MLE solution alone.

We close with some interesting directions for future work. We have seen that LASSO-based variable selection in the presence of implosion breakdown can lead to missed detection of relevant predictors. This motivates the question of whether explosion breakdown can lead to the inclusion of irrelevant predictors. Finally, with respect to convergence issues of our algorithm, while we have established conditions under which our algorithm is guaranteed to converge to a stationary point we do not have rigorous results on the rate at which it does so. As a complement to methods that may be sensitive to the presence of outliers, characterizing the convergence speed of our algorithm has a great deal of practical importance.

SUPPLEMENTAL MATERIALS

Algorithm details, simulation results, proofs, and derivations: The Supplementary Materials includes additional details on the algorithm (e.g. choosing initial starting points, stopping criteria, and choosing regularization parameters), additional results from the estimation experiments in Section 5.1 and variable selection experiments in Section 5.2, proofs for Theorems 4.1 and 4.2, and a derivation of our HHSVM algorithm. (Supplement.pdf)

Code: C and R code used to generate results shown in the article along with relevant data have also been made available. A readme file details how to compile and run the code. The SNP data is not included for confidentiality reasons. (GNU zipped tar file)

ACKNOWLEDGMENTS

The authors thank Christopher Amos for generously allowing them to work with the lung cancer data set. All plots were made using the open source R package ggplot2 (Wickham, 2009). Eric Chi was supported by grant DE-FG02-97ER25308 from the Department of Energy. David Scott was supported in part by grant DMS-09-07491 from the National Science Foundation.

References

Alfons, A., Croux, C., and Gelper, S. (2012), "Sparse least trimmed squares regression," *Annals of Applied Statistics*, to appear.

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayak̃rishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y.-Y., Chen, W. V., Shete, S. a., Spitz, M. R., and Houlston, R. S. (2008), "Genome-wide Association Scan of Tag SNPs Identifies a Susceptibility Locus for Lung Cancer at 15q25.1," *Nature Genetics*, 40, 616–622.

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998), "Robust and Efficient Estimation by Minimising a Density Power Divergence," *Biometrika*, 85, 549–559.

Bianco, A. and Yohai, V. (1996), "Robust Estimation in the Logistic Regression Models," in *Robust Statistics, Data Analysis, and Computer Intensive Methods, Lecture Notes in Statistics*, ed. Rieder, H., New York: Springer-Verlag, vol. 109, pp. 17–34.

Bondell, H. D. (2005), "Minimum Distance Estimation for the Logistic Regression Models," *Biometrika*, 92, 724–731.

Carroll, R. J. and Pederson, S. (1993), "On Robustness in the Logistic Regression Model," *Journal of the Royal Statistical Society, Ser. B*, 55, 693–706.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998), "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, 20, 33–61.

Chi, E. C. (2011), "Parametric Classification and Variable Selection by the Minimum Integrated Squared Error Criterion," Ph.D. thesis, Rice University.

Collins, M., Dasgupta, S., and Schapire, R. (2001), "A Generalization of Principal Component Analysis to the Exponential Family," in *Advances in Neural Information Processing Systems*, vol. 14.

Copas, J. B. (1988), "Binary Regression Models for Contaminated Data," *Journal of the Royal Statistical Society. Series B*, 50, 225–265.

Croux, C., Flandre, C., and Haesbroeck, G. (2002), "The Breakdown Behavior of the Maximum Likelihood Estimator in the Logistic Regression Models," *Statistics & Probability Letters*, 60, 377–386.

Donoho, D. L. and Johnstone, I. M. (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224.

Donoho, D. L. and Liu, R. C. (1988), "The "Automatic" Robustness of Minimum Distance Functionals," *Annals of Statistics*, 16, 552–586.

Frank, A. and Asuncion, A. (2010), "UCI Machine Learning Repository," .

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise coordinate optimization," *Annals of Applied Statistics*, 1, 302–332.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22.

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304.

Hunter, D. and Lange, K. (2004), "A Tutorial on MM Algorithms." *The American Statistician*, 58, 30–38.

Kim, J. and Scott, C. (2008), "Performance Analysis for L_2 Kernel Classification," in *Advances in Neural Information Processing Systems*, vol. 21.

— (2010), "L_2 Kernel Classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32, 1822–1831.

Kolda, T. G. and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500.

Künsch, H. R., Stefanski, L. A., and Carroll, R. J. (1989), "Conditionally Unbiased Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models," *Journal of the American Statistical Association*, 84, 460–466.

Lange, K. (2010), *Numerical Analysis for Statisticians*, Springer.

Lange, K., Hunter, D. R., and Yang, I. (2000), "Optimization Transfer Using Surrogate Objective Functions," *Journal of Computational and Graphical Statistics*, 9, 1–20.

Lee, S., Huang, J. Z., and Hu, J. (2010), "Sparse Logistic Principal Components Analysis for Binary Data," *Annals of Applied Statistics*, 4, 1579–1601.

Li, G., Peng, H., and Zhu, L. (2011), "Nonconcave Penalized M-Estimation with a Diverging Number of Parameters," *Statistica Sinica*, 21, 391–419.

Li, Y., Ding, J., and Abecasis, G. R. (2006), "Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference." *American Journal of Human Genetics*, S79, 2290.

Liu, Z., Jiang, F., Tian, G., Wang, S., Sato, F., Meltzer, S. J., and Tan, M. (2007), "Sparse Logistic Regression with Lp Penalty for Biomarker Identification," *Statistical Applications in Genetics and Molecular Biology*, 6, 2–12.

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, Boca Raton, Florida: Chapman and Hall.

Rosset, S. and Zhu, J. (2007), "Piecewise Linear Regularized Solution Paths," *Annals of Statistics*, 35, 1012–1030.

Schifano, E. D., Strawderman, R. L., and Wells, M. T. (2010), "Majorization-Minimization Algorithms for Nonsmoothly Penalized Objective Functions," *Electronic Journal of Statistics*, 4, 1258–1299.

Scott, D. W. (1992), *Multivariate Density Estimation. Theory, Practice and Visualization*, John Wiley & Sons, Inc.

— (2001), "Parametric Statistical Modeling by Minimum Integrated Square Error," *Technometrics*, 43, 274–285.

— (2004), "Partial Mixture Estimation and Outlier Detection in Data and Regression," in *Theory and Applications of Recent Robust Methods*, eds. Hubert, M., Pison, G., Struyf, A., and Aelst, S. V., Birkhauser, Basel, pp. 297–306.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

van de Geer, S. A. (2008), "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, 36, 614–645.

Wang, H., Li, G., and Jiang, G. (2007), "Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso," *Journal of Business & Economic Statistics*, 25, 347–355.

Wang, L., Zhu, J., and Zou, H. (2008), "Hybrid Huberized Support Vector Machines for Microarray Classification and Gene Selection," *Bioinformatics*, 24, 412–419.

Wickham, H. (2009), *ggplot2: Elegant Graphics for Data Analysis*, Springer New York.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009), "Genomewide Association Analysis by Lasso Penalized Logistic Regression," *Bioinformatics*, 25, 714–721.

Wu, T. T. and Lange, K. (2008), "Coordinate Descent Algorithms for Lasso Penalized Regression," *Annals of Applied Statistics*, 2, 224–244.

Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society: Series B*, 68, 49–67.

Zou, H. and Hastie, T. (2005), “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society, Ser. B*, 67, 301–320.