Supplementary Materials for "Robust Parametric Classification and Variable Selection by a Minimum Distance Criterion"

Eric C. Chi and David W. Scott

Contents

1 Proofs

1.1 Proof of Theorem 4.1

It is immediate that $L(\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\theta}}) = L(\mathbf{y},\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})$. We turn our attention to proving that $L(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) \geq L(\mathbf{y},\tilde{\mathbf{X}}\boldsymbol{\theta})$ for all $\boldsymbol{\theta},\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{p+1}$. Since $L(\mathbf{y},\tilde{\mathbf{X}}\boldsymbol{\theta})$ has bounded curvature our strategy is to represent $L(\mathbf{y},\tilde{\mathbf{X}}\boldsymbol{\theta})$ 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. This approach applies in general to functions with continuous second derivative and bounded curvature (Böhning and Lindsay, 1988).

The exact second order Taylor expansion of $L(\mathbf{y},\tilde{\mathbf{X}}\boldsymbol{\theta})$ at $\tilde{\boldsymbol{\theta}}$ is given by

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

where $\boldsymbol{\theta}^* = \gamma\tilde{\boldsymbol{\theta}} + (1-\gamma)\boldsymbol{\theta}$ for some $\gamma \in (0,1)$ and

$$\begin{aligned}
\nabla L(\mathbf{y},\tilde{\mathbf{X}}\boldsymbol{\theta}) &= 4n^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{G}(\mathbf{p}-\mathbf{y}) \\
\mathbf{H}_{\boldsymbol{\theta}} &= \frac{2}{n}\mathbf{X}^{\mathsf{T}}\mathbf{M}_{\boldsymbol{\theta}}\mathbf{X}, \\
\mathbf{G} &= \mathrm{diag}\{p_1(1-p_1),\ldots,p_n(1-p_n)\} \\
\mathbf{M}_{\boldsymbol{\theta}} &= \mathrm{diag}\{\psi_{u_1}(p_1),\ldots,\psi_{u_n}(p_n)\} \\
\mathbf{u} &= 2\mathbf{y}-\mathbf{1} \\
\mathbf{p} &= F(\tilde{\mathbf{X}}\boldsymbol{\theta}) \\
\psi_u(p) &= [2p(1-p)-(2p-1)((2p-1)-u)]p(1-p).
\end{aligned}$$

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:

$$L(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}}) = L(\mathbf{y},\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) + \frac{4}{n}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{G}(F(\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})-\mathbf{y}) + \frac{\eta}{n}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}),$$

where

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

Note that for any $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$, $(\mathbf{M}_{\boldsymbol{\theta}})_{ii} \leq \eta$. Therefore,

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

and consequently $L(\boldsymbol{\theta};\tilde{\boldsymbol{\theta}})$ majorizes $L(\mathbf{y},\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})$ 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

$$\sup_{p\in[0,1]} \psi_1(p) = \max_{p\in[0,1]} \psi_1(p) = \frac{1}{4} \max_{q\in[-1,1]} \left\{ \frac{3}{2}q^4 - q^3 - 2q^2 + q + \frac{1}{2} \right\}. \tag{1.1}$$

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 $q = 2p-1$. Since the derivative of the polynomial in (1.1) has a root at 1, it is straightforward to argue that the lower bound of η is attained at the second largest root, which is $(-3+\sqrt{33})/12$. Thus, the majorization holds so long as

$$\eta \geq \frac{3}{16}q^4 - \frac{1}{4}q^3 - \frac{1}{2}q^2 + \frac{1}{4}q + \frac{1}{16}\bigg|_{q=\frac{-3+\sqrt{33}}{12}}.$$

1.2 Proof of Theorem 4.2

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

We will use the MM algorithm global convergence results in Schifano et al. (2010). Let $\xi(\boldsymbol{\theta})$ denote the objective function and let $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ denote a surrogate objective function that will be minimized with respect to its first argument in lieu of $\xi(\boldsymbol{\theta})$. The iteration map φ is given by

$$\varphi(\tilde{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\theta}} \xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}).$$

We now state a slightly less general set of regularity conditions than those in Schifano et al. (2010) that are sufficient for our purposes. Suppose $\xi, \xi^{[S]}$, and φ satisfy the following set of conditions:

R1. The objective function $\xi(\boldsymbol{\theta})$ is locally Lipschitz continuous for $\boldsymbol{\theta} \in \Theta$ and coercive. The set of stationary points $\mathcal{S}$ of $\xi(\boldsymbol{\theta})$ is a finite set, where the notion of a stationary point is defined as in Clarke (1983).

R2. $\xi(\boldsymbol{\theta}) = \xi^{[S]}(\boldsymbol{\theta}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$.

R3. $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) < \xi^{[S]}(\boldsymbol{\theta}, \boldsymbol{\theta})$ for all $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta$ where $\boldsymbol{\theta} \neq \tilde{\boldsymbol{\theta}}$.

R4. $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is continuous for $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \in \Theta \times \Theta$ and locally Lipschitz in Θ.

R5. $\varphi(\boldsymbol{\theta})$ is a singleton set consisting of one bounded vector for $\boldsymbol{\theta} \in \Theta$.

Then $\{\boldsymbol{\theta}^{(n)}, n \geq 0\}$ converges to a fixed point of the iteration map φ. By Proposition A.8 in Schifano et al. (2010) the fixed points of φ coincide with $\mathcal{S}$.

In our case we have the following objective and surrogate functions

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

We check each regularity condition in turn.

R1. Since $\|\mathbf{y} - F(\tilde{\mathbf{X}}\boldsymbol{\theta})\|_2^2$ is bounded below and the penalty term is coercive, $\xi(\boldsymbol{\theta})$ is coercive. Recall that the gradient of the $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ is $(4/n)\mathbf{X}^\mathsf{T}\mathbf{G}(F(\tilde{\mathbf{X}}\boldsymbol{\theta}) - \mathbf{y})$. The norm of the gradient is bounded; specifically it is no greater than $2\sigma_1^2$ where σ_1 is the largest singular value of $\mathbf{X}$. Therefore, $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ is Lipschitz continuous and therefore locally Lipschitz continuous. Consequently, $\xi(\boldsymbol{\theta})$ is locally Lipschitz continuous. If the set of stationary points of $\xi(\boldsymbol{\theta})$ is finite, then R1 is met.

R2 and R3. Recall the majorization we are using is given by

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

where

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

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 $L(\mathbf{y}, \tilde{\mathbf{X}}\boldsymbol{\theta})$ and the majorization is strict. R2 and R3 are met.

R4. The penalized majorization is the sum of continuous functions in $(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \in \Theta \times \Theta$ 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. R4 is met.

R5. If $\lambda(1 - \alpha) > 0$ then $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is strictly convex in $\boldsymbol{\theta}$ and thus has at most one global minimizer. Since $\xi^{[S]}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$ is also coercive in $\boldsymbol{\theta}$ it has at least one global minimizer. R5 is met.

Thus, Algorithm 1 will converge to a stationary point of $\xi(\boldsymbol{\theta})$, provided that there are only finitely many stationary points and the coordinate descent minimization of the Elastic Net penalized quadratic majorization is solved exactly. □

Remark 1. *If ξ does not have finitely many stationary points, it can be shown that the limit points of the sequence of iterates are stationary points and that the set of limit points is connected (Schifano et al., 2010; Chi, 2011).*

Remark 2. *The iterate update $\boldsymbol{\theta}^{(m+1)} = \varphi(\boldsymbol{\theta}^{(m)})$ can be accomplished by any means algorithmically so long as the global minimum of the majorization is found. Iterates of coordinate descent are guaranteed to converge to a global minimizer provided that the loss is differentiable and convex and the penalty is convex and separable (Tseng, 2001). Thus, applying coordinate descent on the Elastic Net penalized quadratic majorization will find the global minimum.*

Remark 3. *Our definition of stationary points has to change because the objective functions of interest are locally Lipschitz continuous and therefore differentiable almost everywhere except on a set of Lebesgue measure zero. Clarke (1983) defines and proves properties of a generalized gradient for locally Lipschitz functions. Apart from pathological cases, when a function is convex the generalized gradient is the subdifferential. See Proposition 2.2.7 in Clarke (1983). When a function is differentiable the generalized gradient is the gradient. Thus as would be expected a point $\mathbf{x}$ is a stationary point of a locally Lipschitz function if the function's generalized gradient at $\mathbf{x}$ contains $\mathbf{0}$.*

Algorithm 1 ITERATIVE L_2E SOLVER

$\boldsymbol{\theta} \leftarrow$ initial guess
repeat
  $\mathbf{p} \leftarrow F(\tilde{\mathbf{X}}\boldsymbol{\theta})$
  $\mathbf{G} \leftarrow \text{diag}\{\mathbf{p} * (\mathbf{1} - \mathbf{p})\}$
  $\mathbf{z} \leftarrow 2\mathbf{G}(\mathbf{p} - \mathbf{y})$
  $\boldsymbol{\zeta} \leftarrow \mathbf{X}\boldsymbol{\beta} - \frac{1}{\eta}(\mathbf{z} - \overline{z}\mathbf{1})$
  $\beta_0 \leftarrow \beta_0 - \eta^{-1}\overline{z}$
  repeat
    for $k = 1..p$ **do**
      $\mathbf{r} \leftarrow \boldsymbol{\zeta} - (\mathbf{X}\boldsymbol{\beta} - \beta_k \mathbf{x}_k)$
      $\beta_k \leftarrow S\left(\frac{\eta}{n}\mathbf{x}_k^{\mathsf{T}}\mathbf{r}, \lambda\alpha\right) / \left[\frac{\eta}{n}\|\mathbf{x}_k\|_2^2 + \lambda(1-\alpha)\right]$
    end for
  until convergence
until convergence
return $\boldsymbol{\theta}$

2 Algorithm Details

Algorithm 1 gives pseudocode for the resulting iterative solver for a given pair of parameters α and λ. The symbol $*$ denotes the Hadamard element-wise product. In practice we also use active sets to speed up computations. That is, for a given initial $\boldsymbol{\beta}$, we only update the non-zero coordinates of $\boldsymbol{\beta}$, the active set, until there is little change in the active set parameter estimates. The non-active set parameter estimates are then updated once. If they remain zero, the Karush-Kuhn-Tucker (KKT) conditions have been met and a global minimum of (4.4) has been found. If not, then the active set is expanded to include the coordinates whose KKT conditions have been violated and the process is repeated.

2.1 Choosing the penalty parameters

2.1.1 Warm Starts and Calculating Regularization Paths

We will need to compare the regression coefficients obtained at many values of the penalty parameter λ to perform model selection. Typically we can rapidly calculate regression coefficients for a decreasing sequence of values of λ through warm starts. Namely, a solution to the problem using λ_k as a regularization parameter is used as the initial starting value for the iterative algorithm applied to the subsequent problem using λ_{k+1} as a regularization parameter. The idea is if λ_k and λ_{k+1} are not too far apart, the solutions to their corresponding optimization problems will be close to each other. Thus, the solution of one optimization problem will be a very good initial starting point for the succeeding optimization problem.

For λ sufficiently large, only the intercept term θ_0 will come into the model. The smallest λ^* such that all regression coefficients are shrunk to zero is given by

$$\lambda^* = \frac{2}{n\alpha}\overline{y}(1-\overline{y}) \max_{j=1,\ldots,p} |\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{y}|, \tag{2.1}$$

where $\mathbf{x}_{(j)}$ denotes the jth column of the design matrix $\mathbf{X}$. We compute a grid of λ values equally spaced on a log scale between $\lambda_{\max} = \lambda^*$ and $\lambda_{\min} = \epsilon\lambda_{\max}$ where $\epsilon < 1$. 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.

Moreover, due to the possible multi-modality of the L_2E loss, we recommend computing the regulation paths starting from a smaller regularization parameter and increasing the parameter value until $\lambda_{\max}$. Since we face multi-modality initial starting points can make a significant difference in the answers obtained.

2.1.2 The heuristic for choosing starting values

Since the logistic L_2E loss is not convex, it may have multiple local minima. For the purely LASSO-penalized problem, the KKT condition at a local minimum is

$$\nu_j = |\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{G}(\mathbf{y} - F(\beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta}))| \leq \lambda.$$

Equality is met whenever $\beta_j \neq 0$. Thus, the largest values of ν_j will correspond to a set of covariates which include covariates with non-zero regression coefficients. The leap of faith is that the largest values of ν_j evaluated at the null model will also correspond to a set of covariates which include covariates with non-zero regression coefficients. This idea has been used in a "swindle" rule (Wu et al., 2009) and STRONG rules for discarding variables (Tibshirani, Bien, Friedman, Hastie, Simon, Taylor, and Tibshirani, 2012). In those instances the goal is to solve a smaller optimization problem. In contrast, we initialize starting parameter entries to zero rather than excluding variables with low scores from the optimization problem. Specifically, we do the following: (1) calculate the following scores $z_j = |\mathbf{x}_{(j)}^{\mathsf{T}}\mathbf{G}_0(\mathbf{y} - p\mathbf{1}))|$, where $p = \overline{y}$ the sample mean of $\mathbf{y}$ and $\mathbf{G}_0 = p(1-p)\mathbf{I}$; (2) set $\beta_0^{(0)} = \log(\overline{y}/(1-\overline{y}))$; and (3) set $\beta_j^{(0)} = I(j \in \mathcal{S})$, where $I(\cdot)$ denotes the indicator function and $\mathcal{S} = \{j : z_j \text{ is "large"}\}$.

2.1.3 Robust Cross-Validation

Once we have a set of models computed at different regularization parameter values, we select the model that is optimal with respect to some criterion. We use the following robust 10-fold cross-validation scheme to select the model. After partitioning the data into 10 training and test sets, for each $i = 1, \ldots, 10$ folds we compute regression coefficients $\hat{\boldsymbol{\theta}}^{-i}(\lambda)$ for a sequence of λ's between $\lambda_{\max}$ and $\lambda_{\min}$ holding out the ith test set $\mathcal{S}_i$.

Next we refit the model using the reduced variable set $\mathcal{S}_i^c$, those with nonzero regression coefficients, and refit using logistic L_2E with $\alpha = 0$. This refitting produces less biased estimates. We are adopting the same strategy as LARS-OLS in Efron, Hastie, Johnstone, and Tibshirani (2004). Our framework, however, could adopt a more sophisticated strategy along the lines of the Relaxed LASSO in Meinshausen (2007). Henceforth let $\hat{\boldsymbol{\theta}}^{-i}(\lambda)$ denote the regression coefficients obtained after the second step. Let $d_j^{-i}(\lambda)$ denote the contribution of observation j to the L_2E loss under the model $\hat{\boldsymbol{\theta}}^{-i}(\lambda)$, i.e.,

$$d_j^{-i}(\lambda) = \left(y_j - F(\tilde{\mathbf{x}}_j^{\mathsf{T}}\hat{\boldsymbol{\theta}}^{-i}(\lambda))\right)^2.$$

We use the following criterion to choose λ^*:

$$\lambda^* = \arg\min_{\lambda} \left\{ \underset{i=1,\ldots,10}{\text{median}} \left\{ \underset{j \in \mathcal{S}_i}{\text{median}} \, d_j^{-i}(\lambda) \right\} \right\}.$$

The reason for choosing λ^* in this way is due to a feature of the robust fitting procedure. Good robust models will assign unusually large values of $d_j^{-i}(\lambda)$ to outliers. Thus, the total L_2E loss is an inappropriate measure of the prediction error if influential outliers were present. On the other hand, taking the median, for example, would provide a more unbiased measure of the prediction error regardless of outliers. The final model selected would be the one that minimizes the robust prediction error criterion.

3 Simulation Experiments in Low Dimensions

Tables 3 and 4 provide summary statistics for simulations performed in Section 5.1. The experiments show the unbiasedness of the L_2E compared to the MLE at the price of increased variance. The mse summarizes the bias-variance tradeoff between the two methods.

Table 3: Effect of varying the position of a single outlier from -0.25 to 24.

			MLE			L_2E		
Outlier Position	Coefficient	True Value	mean	std	mse	mean	std	mse
-0.25	β_0	0	-0.002	0.182	0.033	-0.005	0.192	0.037
	β_1	1	1.032	0.434	0.189	1.063	0.480	0.234
	β_2	0.5	0.526	0.424	0.180	0.539	0.463	0.216
	β_3	1	1.047	0.439	0.195	1.079	0.482	0.238
	β_4	2	2.110	0.487	0.249	2.181	0.572	0.359
1.5	β_0	0	-0.024	0.168	0.029	0.002	0.192	0.037
	β_1	1	0.868	0.394	0.173	1.052	0.476	0.229
	β_2	0.5	0.401	0.391	0.162	0.532	0.460	0.212
	β_3	1	0.880	0.396	0.171	1.068	0.478	0.233
	β_4	2	1.860	0.430	0.204	2.160	0.567	0.347
3	β_0	0	-0.022	0.157	0.025	0.002	0.192	0.037
	β_1	1	0.732	0.368	0.207	1.054	0.476	0.229
	β_2	0.5	0.296	0.369	0.178	0.533	0.460	0.212
	β_3	1	0.743	0.368	0.201	1.069	0.478	0.233
	β_4	2	1.662	0.392	0.268	2.163	0.567	0.347
6	β_0	0	-0.020	0.142	0.021	0.002	0.192	0.037
	β_1	1	0.508	0.337	0.356	1.054	0.476	0.229
	β_2	0.5	0.112	0.344	0.268	0.533	0.460	0.212
	β_3	1	0.516	0.334	0.346	1.069	0.478	0.233
	β_4	2	1.350	0.347	0.543	2.163	0.567	0.347
12	β_0	0	-0.018	0.128	0.017	0.002	0.192	0.037
	β_1	1	0.153	0.325	0.823	1.054	0.476	0.229
	β_2	0.5	-0.201	0.336	0.604	0.533	0.460	0.212
	β_3	1	0.158	0.316	0.808	1.069	0.478	0.233
	β_4	2	0.906	0.317	1.297	2.163	0.567	0.347
24	β_0	0	-0.011	0.124	0.016	0.002	0.192	0.037
	β_1	1	-0.088	0.330	1.293	1.054	0.476	0.229
	β_2	0.5	-0.431	0.331	0.975	0.533	0.460	0.212
	β_3	1	-0.086	0.315	1.279	1.069	0.478	0.233
	β_4	2	0.641	0.324	1.952	2.163	0.567	0.347

Table 4: Effect of varying the number of outliers at a fixed location.

			MLE			L_2E		
Number of Outliers	Coefficient	True Value	mean	std	mse	mean	std	mse
0	β_0	0	0.005	0.182	0.033	0.002	0.192	0.037
	β_1	1	1.026	0.433	0.188	1.054	0.476	0.229
	β_2	0.5	0.521	0.422	0.179	0.533	0.460	0.212
	β_3	1	1.041	0.438	0.193	1.069	0.478	0.233
	β_4	2	2.099	0.485	0.245	2.163	0.567	0.347
1	β_0	0	-0.022	0.157	0.025	0.002	0.192	0.037
	β_1	1	0.732	0.368	0.207	1.054	0.476	0.229
	β_2	0.5	0.296	0.369	0.178	0.533	0.460	0.212
	β_3	1	0.743	0.368	0.201	1.069	0.478	0.233
	β_4	2	1.662	0.392	0.268	2.163	0.567	0.347
5	β_0	0	-0.090	0.126	0.024	0.002	0.192	0.037
	β_1	1	0.086	0.320	0.937	1.054	0.476	0.229
	β_2	0.5	-0.263	0.327	0.689	0.533	0.460	0.212
	β_3	1	0.090	0.308	0.922	1.069	0.478	0.233
	β_4	2	0.830	0.312	1.466	2.163	0.567	0.347
10	β_0	0	-0.110	0.124	0.027	0.002	0.192	0.037
	β_1	1	-0.073	0.330	1.261	1.054	0.476	0.229
	β_2	0.5	-0.417	0.333	0.951	0.533	0.460	0.212
	β_3	1	-0.071	0.315	1.246	1.069	0.478	0.233
	β_4	2	0.659	0.323	1.903	2.163	0.567	0.347
15	β_0	0	-0.117	0.124	0.029	0.002	0.192	0.037
	β_1	1	-0.127	0.335	1.382	1.054	0.476	0.229
	β_2	0.5	-0.470	0.338	1.055	0.533	0.460	0.212
	β_3	1	-0.125	0.321	1.367	1.069	0.478	0.233
	β_4	2	0.605	0.328	2.054	2.163	0.567	0.347
20	β_0	0	-0.122	0.124	0.030	0.002	0.192	0.037
	β_1	1	-0.159	0.339	1.457	1.054	0.476	0.229
	β_2	0.5	-0.502	0.342	1.120	0.533	0.460	0.212
	β_3	1	-0.157	0.325	1.443	1.069	0.478	0.233
	β_4	2	0.573	0.332	2.145	2.163	0.567	0.347

4 Variable Selection Experiments in High Dimensions

We show more detailed results for a single replicate for the simulations reported in Section 5.2. Figure 1 shows the robust cross validation curves for the three methods for the replicate. Figure 2

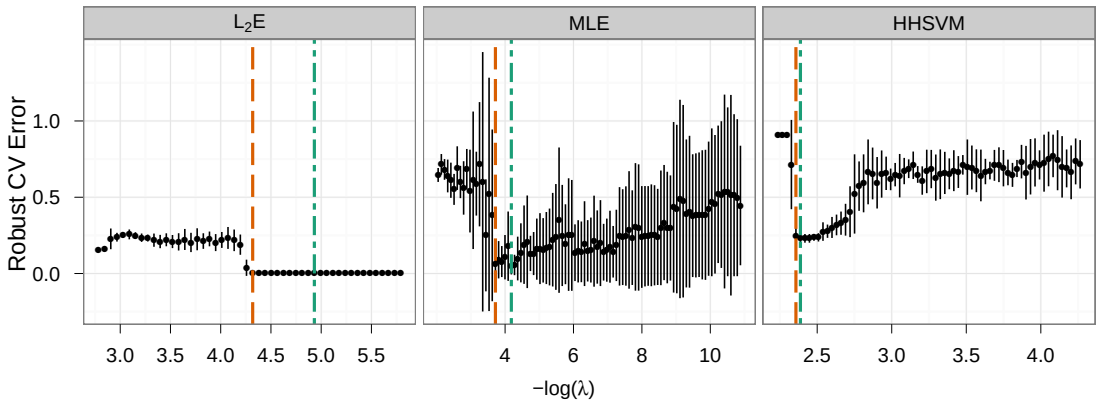

Figure 1: Robust 10-fold cross-validation curves for the three methods. The vertical error bars around the dots indicate ± one median absolute deviation with a scale factor of 1.4826. The dash-dotted line indicates the minimizing λ. The dashed line indicates the 1-MAD rule λ.

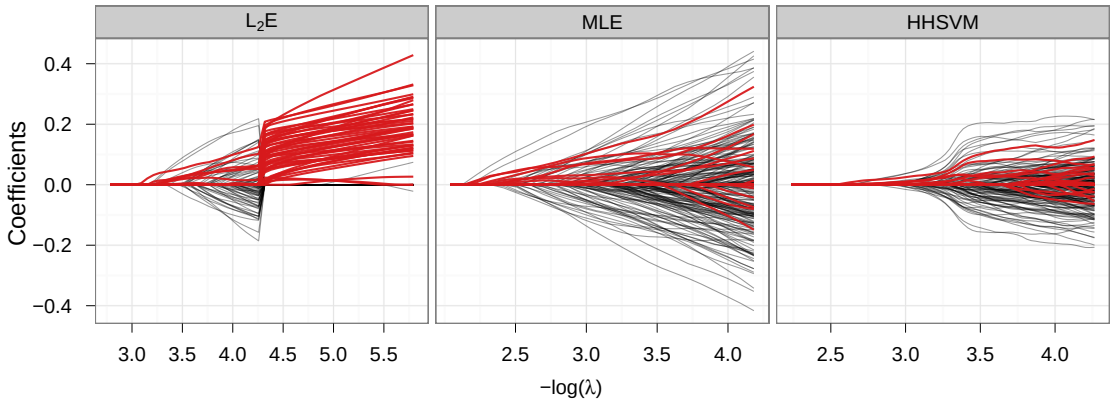

Figure 2: Regularization paths for the three methods. Paths for nonzero regression coefficients in the true model are drawn in heavy solid lines.

shows the regularization paths for the three methods for the replicate. Note the large jump in the L_2E curve. By choosing the starting L_2E point by our heuristic, a local minimum different from the MLE solution is found. For sufficiently large λ, however, the local minimum vanishes, and the regularization paths mimic the MLE regularization paths.

5 The Hybrid Huberized SVM

Consider the following classification problem. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote a centered matrix of covariates and $\mathbf{y} \in \{-1, 1\}^n$ denote binary class labels. We will employ the compact notation $\tilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X}) \in \mathbb{R}^{n \times (p+1)}$ and $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{p+1}$. The Hybrid Huberized Support Vector Machine (HHSVM) (Wang et al., 2008) constructs a linear classifier $\tilde{\mathbf{X}}\boldsymbol{\theta}$ by minimizing the following loss.

$$\ell(\mathbf{y}, \mathbf{X}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \phi\left(y_i \tilde{\mathbf{x}}_i^\mathsf{T} \boldsymbol{\theta}\right) + J(\boldsymbol{\beta}),$$

where the function ϕ is a smooth hinge loss,

$$\phi(u) = \begin{cases} (1-t)^2 + 2(1-t)(t-u), & \text{if } u \le t, \\ (1-u)^2, & \text{if } t < u \le 1, \\ 0, & \text{otherwise,} \end{cases}$$

and J is the Elastic Net penalty (Zou and Hastie, 2005).

$$J(\boldsymbol{\beta}) = \lambda \left(\alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 \right),$$

where $\alpha \in [0, 1]$ is a mixing parameter between the 1-norm and 2-norm regularizers. We now derive an MM algorithm for solving the entire regularization path with respect to a varying λ for a fixed α. The majorization we will use leads to a simple MM algorithm. This algorithm calculates a different regularization path than the algorithm in (Wang et al., 2008), which uses the following parameterization of the Elastic Net

$$J(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2,$$

for varying λ_1 for a fixed λ_2. The code used in (Wang et al., 2008) is available on the author's website (`http://www.stat.lsa.umich.edu/~jizhu/code/hhsvm`).

5.1 An MM Algorithm for Minimizing the Smooth Hinge Loss

We begin by deriving a quadratic majorization of ϕ. It is straightforward to verify that the first and second derivatives of ϕ are given by

$$\phi'(u) = \begin{cases} -2(1-t), & \text{if } u \leq t, \\ -2(1-u), & \text{if } t < u \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

$$\phi''(u) = \begin{cases} 0, & \text{if } u \leq t, \\ 2, & \text{if } t < u \leq 1, \\ 0, & \text{otherwise.} \end{cases}$$

Then we can express ϕ as an exact second order Taylor expansion at a point $\tilde{u}$ with

$$\phi(u) = \phi(\tilde{u}) + \phi'(\tilde{u})(u - \tilde{u}) + \frac{1}{2}\phi''(u^*)(u - \tilde{u})^2,$$

where $u^* = \delta u + (1-\delta)\tilde{u}$ for some $\delta \in (0,1)$. It follows immediately that the following function majorizes ϕ at $\tilde{u}$.

$$g(u; \tilde{u}) = \phi(\tilde{u}) + \phi'(\tilde{u})(u - \tilde{u}) + (u - \tilde{u})^2.$$

The u that minimizes $g(u; \tilde{u})$ is

$$\begin{aligned} u &= \tilde{u} - \frac{1}{2}\phi'(\tilde{u}) \\ &= \tilde{u} + [(1-t)I(u \leq t) + (1-u)I(u > t)I(u \leq 1)] \\ &= \tilde{u} + 1 - \min(\max(\tilde{u}, t), 1) \end{aligned}$$

5.2 An MM Algorithm for the Unregularized Classification Problem

Returning to our original problem and applying the above results along with the chain rule gives us the relationship

$$\ell(\mathbf{y}, \tilde{\mathbf{X}}; \boldsymbol{\theta}) \leq \ell(\mathbf{y}, \tilde{\mathbf{X}}; \tilde{\boldsymbol{\theta}}) + \tilde{\boldsymbol{\varphi}}^\mathsf{T} \tilde{\mathbf{X}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \|\tilde{\mathbf{X}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\|_2^2,$$

where

$$\tilde{\boldsymbol{\varphi}}_i = y_i \varphi'(y_i \tilde{\mathbf{x}}_i^\mathsf{T} \tilde{\boldsymbol{\theta}}).$$

Since the equality occurs when $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$, the right hand side majorizes the left hand side. Furthermore, the majorization up to an additive constant is separable in β_0 and $\boldsymbol{\beta}$.

$$\begin{aligned} \left\| \frac{1}{2}\tilde{\boldsymbol{\varphi}} + \tilde{\mathbf{X}}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \right\|_2^2 &= \left\| (\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}} - \frac{1}{2}\tilde{\boldsymbol{\varphi}}) - \tilde{\mathbf{X}}\boldsymbol{\theta} \right\|_2^2 \\ &= \left\| \left[\mathbf{X}\tilde{\boldsymbol{\beta}} - \frac{1}{2}(\tilde{\boldsymbol{\varphi}} - \overline{\varphi}\mathbf{1}) - \mathbf{X}\boldsymbol{\beta} \right] + \left[\tilde{\beta}_0 \mathbf{1} - \frac{1}{2}\overline{\varphi}\mathbf{1} - \beta_0 \mathbf{1} \right] \right\|_2^2 \\ &= n \left(\tilde{\beta}_0 - \beta_0 - \frac{1}{2n}\mathbf{1}^\mathsf{T}\tilde{\boldsymbol{\varphi}} \right)^2 + \|\tilde{\mathbf{z}} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \end{aligned}$$

where

$$\tilde{\mathbf{z}} = \mathbf{X}\tilde{\boldsymbol{\beta}} - \frac{1}{2}\left(\tilde{\boldsymbol{\varphi}} - \frac{1}{n}\mathbf{1}^{\mathsf{T}}\tilde{\boldsymbol{\varphi}}\mathbf{1}\right).$$

We can write the updates with the intercept and regression coefficients separately. The intercept update is

$$\beta_0 = \tilde{\beta}_0 - \frac{1}{2n}\mathbf{1}^{\mathsf{T}}\tilde{\boldsymbol{\varphi}}.$$

and if $\mathbf{X}$ is full rank the update for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} - \frac{1}{2}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\left(\tilde{\boldsymbol{\varphi}} - \frac{1}{n}\mathbf{1}^{\mathsf{T}}\tilde{\boldsymbol{\varphi}}\mathbf{1}\right).$$

5.3 An MM Algorithm for the HHSVM

Adding an Elastic Net penalty to the majorization gives us the following loss function to minimize.

$$\frac{1}{2}\left(\tilde{\beta}_0 - \beta_0 - \frac{1}{2n}\mathbf{1}^{\mathsf{T}}\tilde{\boldsymbol{\varphi}}\right)^2 + \frac{1}{2n}\left\|\tilde{\mathbf{z}} - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2}\|\boldsymbol{\beta}\|_2^2\right).$$

Penalized least squares problems of this variety are efficiently solved with coordinate descent. The coordinate descent updates are

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

where

$$\mathbf{r} = \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}} - \frac{1}{2}\tilde{\boldsymbol{\varphi}} - \sum_{j \neq k}\beta_j\mathbf{x}_j.$$

References

Böhning, D. and Lindsay, B. G. (1988), "Monotonicity of Quadratic-Approximation Algorithms," *Annals of the Institute of Statistical Mathematics*, 40, 641–663.

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

Clarke, F. H. (1983), *Optimization and Nonsmooth Analysis*, Wiley-Interscience.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics and Data Analysis*, 52, 374–393.

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.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. J. (2012), "Strong rules for discarding predictors in lasso-type problems," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 245–266.

Tseng, P. (2001), "Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization," *Journal of Optimization Theory and Applications*, 109, 475–494.

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

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.

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.