SparseNet: Coordinate Descent with Non-Convex Penalties

Rahul Mazumder Jerome H. Friedman Trevor Hastie *

Abstract

We address the problem of sparse selection in linear models. A number of non-convex penalties have been proposed in the literature for this purpose, along with a variety of convex-relaxation algorithms for finding good solutions. In this paper we pursue a coordinate-descent approach for optimization, and study its convergence properties. We characterize the properties of penalties suitable for this approach, study their corresponding threshold functions, and describe a *df*-standardizing reparametrization that assists our pathwise algorithm. The MC+ penalty (Zhang 2010) is ideally suited to this task, and we use it to demonstrate the performance of our algorithm. Certain technical derivations and experiments of this article are included in the Supplementary Materials Section.

KEYWORDS: Sparse Regression, Regularization Surface, Non-convex Optimization, Coordinate Descent, Degrees of Freedom, Variable Selection, LASSO.

*Rahul Mazumder (E-mail: rahulm@stanford.edu) is a PhD student at Department of Statistics, Stanford University. Jerome H. Friedman (E-mail: jhf@stanford.edu) is Professor Emeritus at Department of Statistics, Stanford University. Trevor Hastie (Email: hastie@stanford.edu) is a Professor at the Departments of Statistics and Health, Research and Policy, Stanford University, Stanford, CA-94305. Trevor Hastie was partially supported by grant DMS-1007719 from the National Science Foundation, and grant RO1-EB001988-12 from the National Institutes of Health. The authors will like to thank the Associate Editor and two referees for valuable comments and suggestions that helped to improve and shorten this presentation.

1 Introduction

Consider the usual linear regression set-up

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

with n observations and p features $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, response $\mathbf{y}$, coefficient vector $\boldsymbol{\beta}$ and (stochastic) error $\boldsymbol{\epsilon}$. In many modern statistical applications with $p \gg n$, the true $\boldsymbol{\beta}$ vector is often sparse, with many redundant predictors having coefficient zero. We would like to identify the useful predictors and also obtain good estimates of their coefficients. Identifying a set of relevant features from a list of many thousand is in general combinatorially hard and statistically troublesome. In this context, convex relaxation techniques such as the LASSO (Tibshirani 1996, Chen & Donoho 1994) have been effectively used for simultaneously producing accurate and parsimonious models. The LASSO solves

$$\min_{\boldsymbol{\beta}} \quad \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{2}$$

The ℓ_1 penalty shrinks coefficients towards zero, and can also set many coefficients to be exactly zero. In the context of variable selection, the LASSO is often thought of as a convex surrogate for best-subset selection:

$$\min_{\boldsymbol{\beta}} \quad \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_0. \tag{3}$$

The ℓ_0 penalty $\|\boldsymbol{\beta}\|_0 = \sum_{i=1}^{p} \mathbf{I}(|\beta_i| > 0)$ penalizes the number of non-zero coefficients in the model.

The LASSO enjoys attractive statistical properties (Zhao & Yu 2006, Donoho 2006, Knight & Fu 2000, Meinshausen & Bühlmann 2006). Under certain regularity conditions on $\mathbf{X}$, it produces models with good prediction accuracy when the underlying model is reasonably sparse. Zhao & Yu (2006) established that the LASSO is

model selection consistent: $\Pr(\hat{\mathcal{A}} = \mathcal{A}^0) \to 1$, where $\mathcal{A}^0$ corresponds to the set of nonzero coefficients (active) in the true model and $\hat{\mathcal{A}}$ those recovered by the LASSO. Typical assumptions limit the pairwise correlations between the variables.

However, when these regularity conditions are violated, the LASSO can be sub-optimal in model selection (Zhang 2010, Zhang & Huang 2008, Friedman 2008, Zou & Li 2008, Zou 2006). Since the LASSO both shrinks and selects, it often selects a model which is overly dense in its effort to relax the penalty on the relevant coefficients. Typically in such situations greedier methods like subset regression and the non-convex methods we discuss here achieve sparser models than the LASSO for the same or better prediction accuracy, and enjoy superior variable-selection properties.

There are computationally attractive algorithms for the LASSO. The piecewise-linear LASSO coefficient paths can be computed efficiently via the LARS (homotopy) algorithm (Efron et al. 2004, Osborne et al. 2000). Coordinate-wise optimization algorithms (Friedman et al. 2009) appear to be the fastest for computing the regularization paths for a variety of loss functions, and scale well. One-at-a time coordinate-wise methods for the LASSO make repeated use of the univariate *soft-thresholding* operator

$$\begin{aligned} \mathrm{S}(\tilde{\beta}, \lambda) &= \arg\min_{\beta} \left\{ \tfrac{1}{2}(\beta - \tilde{\beta})^2 + \lambda|\beta| \right\} \\ &= \mathrm{sgn}(\tilde{\beta})(|\tilde{\beta}| - \lambda)_+. \end{aligned} \tag{4}$$

In solving (2), the one-at-a-time coordinate-wise updates are given by

$$\tilde{\beta}_j = \mathrm{S}\left(\sum_{i=1}^{n} (y_i - \tilde{y}_i^j) x_{ij}, \lambda \right), \quad j = 1, \ldots, p \tag{5}$$

where $\tilde{y}_i^j = \sum_{k \neq j} x_{ik} \tilde{\beta}_k$ (assuming each $\mathbf{x}_j$ is standardized to have mean zero and unit ℓ_2 norm). Starting with an initial guess for $\tilde{\boldsymbol{\beta}}$ (typically a solution at the previous value for λ), we cyclically update the parameters using (5) until convergence.

The LASSO can fail as a variable selector. In order to get the full effect of a

relevant variable, we have to relax the penalty, which lets in other redundant but possibly correlated features. This is in contrast to best-subset regression; once a strong variable is included and fully fit, it drains the effect of its correlated surrogates.

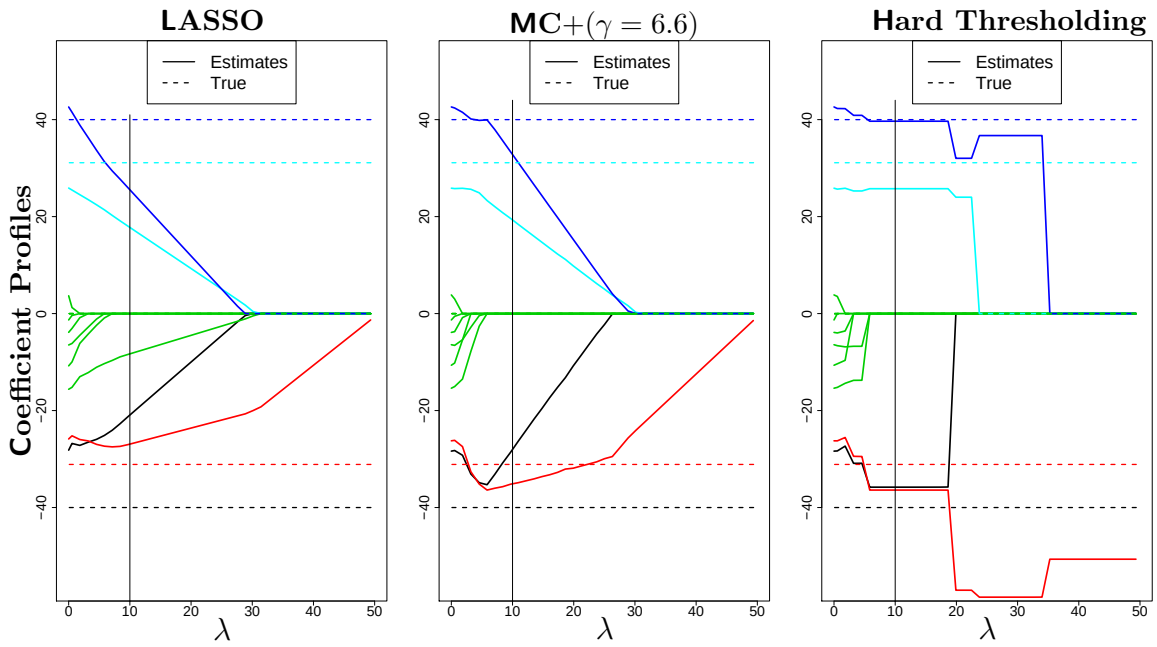

Figure 1: *Regularization path for LASSO (left), MC+ (non-convex) penalized least-squares for* $\gamma = 6.6$ *(centre) and* $\gamma = 1+$ *(right), corresponding to the hard-thresholding operator (best subset). The true coefficients are shown as horizontal dotted lines. Here LASSO is sub-optimal for model selection, as it can never recover the true model. The other two penalized criteria are able to select the correct model (vertical lines), with the middle one having smoother coefficient profiles than best subset on the right.*

As an illustration, Figure 1 shows the regularization path of the LASSO coefficients for a situation where it is sub-optimal for model selection (The simulation setup is defined in Section 7. Here $n = 40$, $p = 10$, SNR$= 3$, $\boldsymbol{\beta} = (-40, -31, \mathbf{0}_{1\times 6}, 31, 40)$ and $X \sim \text{MVN}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{diag}[\boldsymbol{\Sigma}(0.65; 5), \boldsymbol{\Sigma}(0; 5)]$.)

This motivates going beyond the ℓ_1 regime to more aggressive *non-convex* penalties (see the left-hand plots for each of the four penalty families in Figure 2), bridging the gap between ℓ_1 and ℓ_0 (Fan & Li 2001, Zou 2006, Zou & Li 2008, Friedman 2008,

Zhang 2010). Similar to (2), we minimize

$$Q(\boldsymbol{\beta}) = \quad \tfrac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{p} P(|\beta_i|; \lambda; \gamma), \tag{6}$$

where $P(|\beta|; \lambda; \gamma)$ defines a family of penalty functions concave in $|\beta|$, and λ and γ control the degrees of regularization and concavity of the penalty respectively.

The main challenge here is in the minimization of the possibly non-convex objective $Q(\boldsymbol{\beta})$. As one moves down the continuum of penalties from ℓ_1 to ℓ_0, the optimization potentially becomes combinatorially hard (the optimization problems become non-convex when the non-convexity of the penalty is no longer dominated by the convexity of the squared error loss) .

Our contributions in this paper are as follows.

1. We propose a coordinate-wise optimization algorithm *SparseNet* for finding minima of $Q(\boldsymbol{\beta})$. Our algorithm cycles through both λ and γ, producing solution surfaces $\hat{\boldsymbol{\beta}}_{\lambda,\gamma}$ for all families simultaneously. For each value of λ, we start at the LASSO solution, and then update the solutions via coordinate descent as γ changes, moving us towards best-subset regression.

2. We study the generalized univariate thresholding functions $S_\gamma(\tilde{\beta}, \lambda)$ that arise from different non-convex penalties (6), and map out a set of properties that make them more suitable for coordinate descent. In particular, we seek continuity (in $\tilde{\beta}$) of these threshold functions for both λ and γ.

3. We prove *convergence* of coordinate descent for a useful subclass of non-convex penalties, generalizing the results of Tseng & Yun (2009) to nonconvex problems. Our results go beyond those of Tseng (2001) and Zou & Li (2008); they study stationarity properties of *limit points*, and not convergence of the sequence produced by the algorithms.

4. We propose a re-parametrization of the penalty families that makes them even

more suitable for coordinate-descent. Our re-parametrization constrains the coordinate-wise *effective degrees of freedom* at any value of λ to be constant as γ varies. This in turn allows for a natural transition across neighboring solutions as as we move through values of γ from the convex LASSO towards best-subset selection, with the size of the active set decreasing along the way.

5. We compare our algorithm to the state of the art for this class of problems, and show how our approaches lead to improvements.

Note that this paper is about an algorithm for solving a non-convex optimization problem. What we produce is a good estimate for the solution surfaces. We do not go into methods for selecting the tuning paramaters, nor the properties of the resulting estimators.

The paper is organized as follows. In Section 2 we study four families of non-convex penalties, and their induced thresholding operators. We study their properties, particularly from the point of view of coordinate descent. We propose a degree of freedom (df) calibration, and lay out a list of desirable properties of penalties for our purposes. In Section 3 we describe our *SparseNet* algorithm for finding a surface of solutions for all values of the tuning parameters. In Section 5 we illustrate and implement our approach using the MC+ penalty (Zhang 2010) or the firm shrinkage threshold operator (Gao & Bruce 1997). In Section 6 we study the convergence properties of our *SparseNet* algorithm. Section 7 presents simulations under a variety of conditions to demonstrate the performance of *SparseNet*. Section 8 investigates other approaches, and makes comparisons with *SparseNet* and multi-stage Local Linear Approximation (MLLA/LLA) (Zou & Li 2008, Zhang 2009, Candes et al. 2008). The proofs of lemmas and theorems are gathered in the appendix.

2 Generalized Thresholding Operators

In this section we examine the one-dimensional optimization problems that arise in coordinate descent minimization of $Q(\boldsymbol{\beta})$. With squared-error loss, this reduces to minimizing

$$Q^{(1)}(\beta) = \tfrac{1}{2}(\beta - \tilde{\beta})^2 + \lambda P(|\beta|, \lambda, \gamma). \tag{7}$$

We study (7) for different non-convex penalties, as well as the associated *generalized threshold operator*

$$\mathrm{S}_\gamma(\tilde{\beta}, \lambda) = \arg\min_\beta Q^{(1)}(\beta). \tag{8}$$

As γ varies, this generates a family of threshold operators $\mathrm{S}_\gamma(\cdot, \lambda) : \Re \to \Re$. The soft-threshold operator (4) of the LASSO is a member of this family. The hard-thresholding operator (9) can also be represented in this form

$$\begin{aligned} \mathrm{H}(\tilde{\beta}, \lambda) &= \arg\min_\beta \left\{ \tfrac{1}{2}(\beta - \tilde{\beta})^2 + \lambda\, \mathbf{I}(|\beta| > 0) \right\} \\ &= \tilde{\beta}\, \mathbf{I}\,(|\tilde{\beta}| \geq \lambda). \end{aligned} \tag{9}$$

Our interest in thresholding operators arose from the work of She (2009), who also uses them in the context of sparse variable selection, and studies their properties for this purpose. Our approaches differ, however, in that our implementation uses coordinate descent, and exploits the structure of the problem in this context.

For a better understanding of non-convex penalties and the associated threshold operators, it is helpful to look at some examples. For each penalty family (a)–(d), there is a pair of plots in Figure 2; the left plot is the penalty function, the right plot the induced thresholding operator.

(a) The ℓ_γ penalty given by $\lambda P(t; \lambda; \gamma) = \lambda |t|^\gamma$ for $\gamma \in [0, 1]$, also referred to as the bridge or power family (Frank & Friedman 1993, Friedman 2008).

(b) The log-penalty is a generalization of the elastic net family (Friedman 2008) to

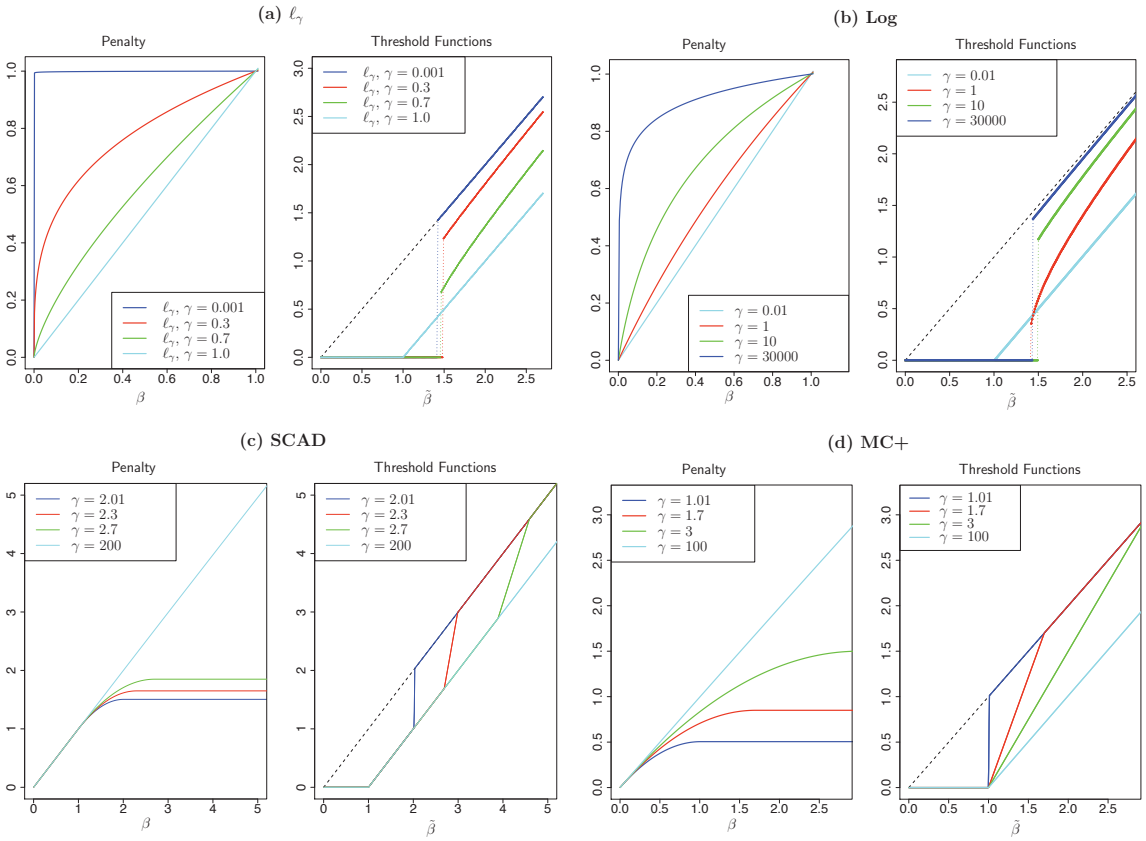

Figure 2: *Non-Convex penalty families and their corresponding threshold functions. All are shown with* $\lambda = 1$ *and different values for* γ.

cover the non-convex penalties from LASSO down to best subset.

$$\lambda P(t;\lambda;\gamma) = \frac{\lambda}{\log(\gamma+1)} \log(\gamma|t|+1), \gamma > 0, \tag{10}$$

where for each value of λ we get the entire continuum of penalties from ℓ_1 ($\gamma \to 0+$) to ℓ_0 ($\gamma \to \infty$).

(c) The SCAD penalty (Fan & Li 2001) is defined via

$$\begin{aligned} \frac{d}{dt}P(t;\lambda;\gamma) &= \mathbf{I}(t \le \lambda) + \frac{(\gamma\lambda - t)_+}{(\gamma-1)\lambda}\mathbf{I}(t > \lambda) \text{ for } t > 0, \quad \gamma > 2 \\ P(t;\lambda;\gamma) &= P(-t;\lambda;\gamma) \\ P(0;\lambda;\gamma) &= 0 \end{aligned} \tag{11}$$

(d) The MC+ family of penalties (Zhang 2010) is defined by

$$\begin{aligned}\lambda P(t;\lambda;\gamma) &= \lambda \int_0^{|t|} (1 - \frac{x}{\gamma\lambda})_+ \, dx \\ &= \lambda(|t| - \frac{t^2}{2\lambda\gamma}) \, \mathbf{I}(|t| < \lambda\gamma) + \frac{\lambda^2\gamma}{2} \, \mathbf{I}(|t| \geq \lambda\gamma). \qquad (12)\end{aligned}$$

For each value of $\lambda > 0$ there is a continuum of penalties and threshold operators, varying from $\gamma \to \infty$ (soft threshold operator) to $\gamma \to 1+$ (hard threshold operator). The MC+ is a reparametrization of the *firm shrinkage* operator introduced by Gao & Bruce (1997) in the context of wavelet shrinkage.

Other examples of non-convex penalties include the transformed ℓ_1 penalty (Nikolova 2000) and the clipped ℓ_1 penalty (Zhang 2009).

Although each of these four families bridge ℓ_1 and ℓ_0, they have different properties. The two in the top row in Figure 2, for example, have discontinuous univariate threshold functions, which would cause instability in coordinate descent. The threshold operators for the ℓ_γ, log-penalty and the MC+ form a continuum between the soft and hard-thresholding functions. The family of SCAD threshold operators, although continuous, do not include $\mathrm{H}(\cdot, \lambda)$. We study some of these properties in more detail in Section 4.

3 *SparseNet*: Algorithm to Construct the Regularization Surface $\hat{\boldsymbol{\beta}}_{\lambda,\gamma}$

We now present our *SparseNet* algorithm for obtaining a family of solutions $\hat{\boldsymbol{\beta}}_{\gamma,\lambda}$ to (6). The $\mathbf{X}$ matrix is assumed to be standardized with each column having zero mean and unit ℓ_2 norm. For simplicity, we assume $\gamma = \infty$ corresponds to the LASSO and $\gamma = 1+$, the hard-thresholding members of the penalty families. The basic idea is as follows. For $\gamma = \infty$, we compute the exact solution path for $Q(\boldsymbol{\beta})$ as a function of λ using coordinate-descent. These solutions are used as warm-starts for

the minimization of $Q(\boldsymbol{\beta})$ at a smaller value of γ, corresponding to a more non-convex penalty. We continue in this fashion, decreasing γ, till we have the solutions paths across a grid of values for γ. The details are given in Algorithm 1.

Algorithm 1 *SparseNet*

1. Input a grid of increasing λ values $\Lambda = \{\lambda_1, \ldots, \lambda_L\}$, and a grid of increasing γ values $\Gamma = \{\gamma_1, \ldots, \gamma_K\}$, where γ_K indexes the LASSO penalty. Define λ_{L+1} such that $\hat{\boldsymbol{\beta}}_{\gamma_K, \lambda_{L+1}} = 0$.
2. For each value of $\ell \in \{L,\ L-1, \ldots, 1\}$ repeat the following
   (a) Initialize $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\gamma_K, \lambda_{\ell+1}}$.
   (b) For each value of $k \in \{K, K-1, \ldots, 1\}$ repeat the following
      i. Cycle through the following one-at-a-time updates $j = 1, \ldots, p, 1, \ldots, p, \ldots$

$$\tilde{\beta}_j = \mathrm{S}_{\gamma_k}\left(\sum_{i=1}^{n}(y_i - \tilde{y}_i^j)x_{ij}, \lambda_\ell\right), \tag{13}$$

      where $\tilde{y}_i^j = \sum_{k \neq j} x_{ik}\tilde{\beta}_k$, until the updates converge to $\boldsymbol{\beta}^*$.
      ii. Assign $\hat{\boldsymbol{\beta}}_{\gamma_k, \lambda_\ell} \leftarrow \boldsymbol{\beta}^*$.
   (c) Decrement k.
3. Decrement ℓ.
4. Return the two-dimensional solution surface $\hat{\boldsymbol{\beta}}_{\lambda,\gamma}$, $(\lambda, \gamma) \in \Lambda \times \Gamma$

In Section 4 we discuss certain properties of penalty functions and their threshold functions that are suited to this algorithm. We also discuss a particular form of *df* recalibration that provides attractive warm-starts and at the same time adds statistical meaning to the scheme.

We have found two variants of Algorithm 1 useful in practice:

- In computing the solution at (γ_k, λ_ℓ), the algorithm uses as a warm start the solution $\hat{\boldsymbol{\beta}}_{\gamma_{k+1}, \lambda_\ell}$. We run a parallel coordinate descent with warm start $\hat{\boldsymbol{\beta}}_{\gamma_k, \lambda_{\ell+1}}$, and then pick the solution with a smaller value for the objective function. This often leads to improved and smoother objective-value surfaces.

- It is sometimes convenient to have a different λ sequence for each value of γ—for example, our recalibrated penalties lead to such a scheme using a doubly indexed sequence $\lambda_{k\ell}$.

In Section 5 we implement *SparseNet* using the calibrated MC+ penalty, which enjoys all the properties outlined in Section 4.3 below. We show that the algorithm converges (Section 6) to a stationary point of $Q(\boldsymbol{\beta})$ for every (λ, γ).

4 Properties of Families of Non-Convex Penalties

Not all non-convex penalties are suitable for use with coordinate descent. Here we describe some desirable properties, and a recalibration suitable for our optimization Algorithm 1.

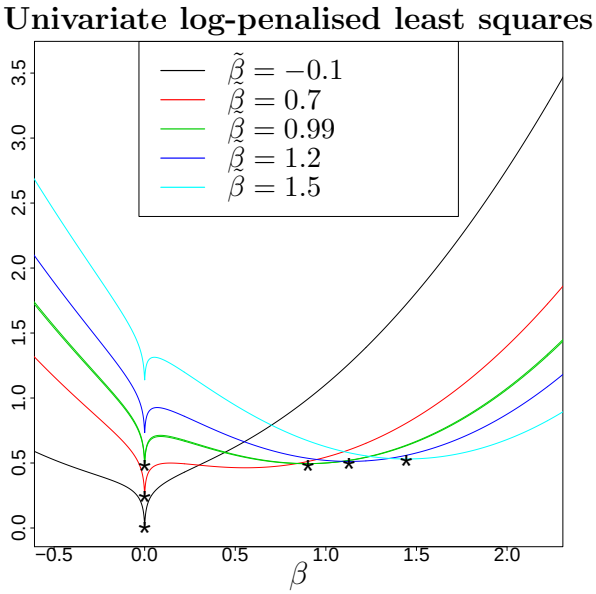

Figure 3: *The penalized least-squares criterion (8) with the log-penalty (10) for* $(\gamma, \lambda) = (500, 0.5)$ *for different values of* $\tilde{\beta}$. *The "*" denotes the global minima of the functions. The "transition" of the minimizers, creates discontinuity in the induced threshold operators.*

4.1 Effect of Multiple Minima in Univariate Criteria

Figure 3 shows the univariate penalized criterion (7) for the log-penalty (10) for certain choices of $\tilde{\beta}$ and a (λ, γ) combination. Here we see the non-convexity of $Q^1(\beta)$, and the transition (with $\tilde{\beta}$) of the global minimizers of the univariate functions. This causes the discontinuity of the induced threshold operators $\tilde{\beta} \mapsto \mathrm{S}_\gamma(\tilde{\beta}, \lambda)$, as shown in Figure 2(b). Multiple minima and consequent discontinuity appears in the $\ell_\gamma, \gamma < 1$ penalty as well, as seen in Figure 2(a).

It has been observed (Breiman 1996, Fan & Li 2001) that discontinuity of the threshold operators leads to increased variance (and hence poor risk properties) of the estimates. We observe that for the log-penalty (for example), coordinate-descent can produce multiple *limit points* (without converging) — creating statistical instability in the optimization procedure. We believe discontinuities such as this will naturally affect other optimization algorithms; for example those based on sequential convex relaxation such as MLLA. In the Supplementary Materials Section 1.4 we see that MLLA gets stuck in a suboptimal local minimum, even for the univariate log-penalty problem. This phenomenon is aggravated for the ℓ_γ penalty.

Multiple minima and hence the discontinuity problem is not an issue in the case of the MC+ penalty or the SCAD (Figures 2(c,d)).

Our study of these phenomena leads us to conclude that if the *univariate* functions $Q^{(1)}(\beta)$ (7) are strictly convex, then the coordinate-wise procedure is well-behaved and converges to a stationary point. This turns out to be the case for the MC+ for $\gamma > 1$ and the SCAD penalties. Furthermore we see in Section 5 in (18) that this restriction on the MC+ penalty for γ still gives us the entire continuum of threshold operators from soft to hard thresholding.

Strict convexity of $Q^{(1)}(\beta)$ also occurs for the log-penalty for some choices of λ and γ, but not enough to cover the whole family. For example, the ℓ_γ family with $\gamma < 1$ does not qualify. This is not surprising since with $\gamma < 1$ it has an unbounded derivative at zero, and is well known to be unstable in optimization.

4.2 Effective *df* and Nesting of *shrinkage-thresholds*

For a LASSO fit $\hat{\mu}_\lambda(x)$, λ controls the extent to which we (over)fit the data. This can be expressed in terms of the *effective degrees of freedom* of the estimator. For our non-convex penalty families, the *df* are influenced by γ as well. We propose to recalibrate our penalty families so that for fixed λ, the coordinate-wise *df* do not change with γ. This has important consequences for our pathwise optimization strategy.

For a linear model, *df* is simply the number of parameters fit. More generally, under an additive error model, *df* is defined by (Stein 1981, Efron et al. 2004)

$$df(\hat{\mu}_{\lambda,\gamma}) = \sum_{i=1}^{n} \mathrm{Cov}(\hat{\mu}_{\lambda,\gamma}(x_i), y_i)/\sigma^2, \tag{14}$$

where $\{(x_i, y_i)\}_1^n$ is the training sample and σ^2 is the noise variance. For a LASSO fit, the *df* is estimated by the number of nonzero coefficients (Zou et al. 2007). Suppose we compare the LASSO fit with k nonzero coefficients to an unrestricted least-squares fit in those k variables. The LASSO coefficients are shrunken towards zero, yet have the same *df*? The reason is the LASSO is "charged" for identifying the nonzero variables. Likewise a model with k coefficients chosen by best-subset regression has *df* greater than k (here we do not have exact formulas). Unlike the LASSO, these are not shrunk, but are being charged the extra *df* for the search.

Hence for both the LASSO and best subset regression we can think of the effective *df* as a function of λ. More generally, penalties corresponding to a smaller degree of non-convexity shrink more than those with a higher degree of non-convexity, and are hence charged less in *df* per nonzero coefficient. For the family of penalized regressions from ℓ_1 to ℓ_0, *df* is controlled by both λ and γ.

In this paper we provide a re-calibration of the family of penalties $P(\cdot; \lambda; \gamma)$ such that for every value of λ the coordinate-wise *df* across the entire continuum of γ values are approximately the same. Since we rely on coordinate-wise updates in the optimization procedures, we will ensure this by calibrating the *df* in the univariate

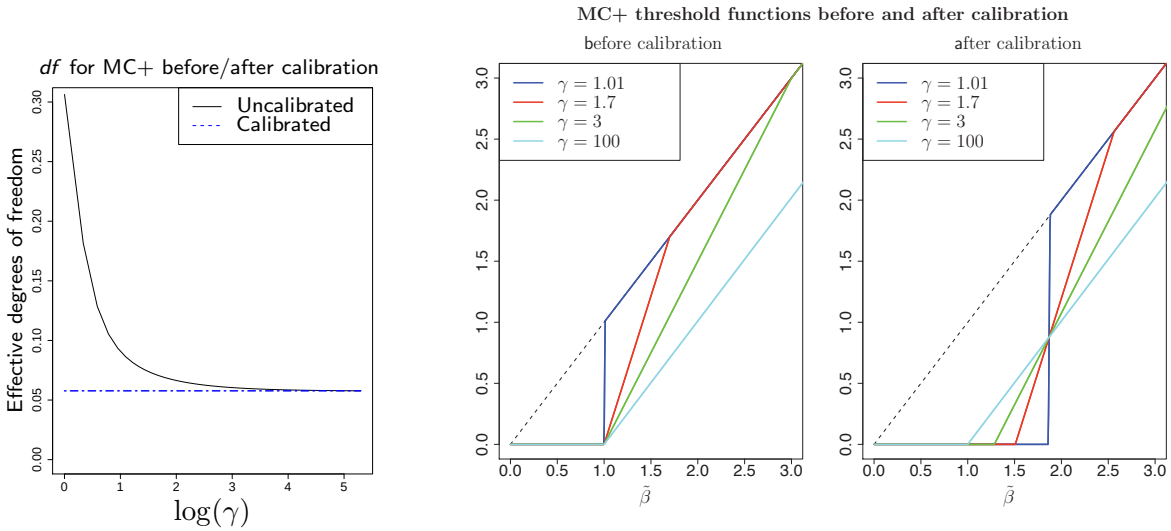

Figure 4: *[Left] The* df *(solid line) for the (uncalibrated) MC+ threshold operators as a function of* γ*, for a fixed* $\lambda = 1$*. The dotted line shows the* df *after calibration. [Middle] Family of MC+ threshold functions, for different values of* γ*, before calibration. All have shrinkage threshold* $\lambda_S = 1$*. [Right] Calibrated versions of the same. The* shrinkage threshold *of the soft-thresholding operator is* $\lambda = 1$*, but as* γ *decreases,* λ_S *increases, forming a continuum between soft and hard thresholding.*

thresholding operators. Details are given in Section 5.1 for a specific example.

We define the *shrinkage-threshold* λ_S of a thresholding operator as the largest (absolute) value that is set to zero. For the soft-thresholding operator of the LASSO, this is λ itself. However, the LASSO also shrinks the non-zero values toward zero. The hard-thresholding operator, on the other hand, leaves all values that survive its shrinkage threshold λ_H alone; in other words it fits the data more aggressively. So for the *df* of the soft and hard thresholding operators to be the same, the *shrinkage-threshold* λ_H for hard thresholding should be larger than the λ for soft thresholding. More generally, to maintain a constant *df* there should be a monotonicity (increase) in the *shrinkage-thresholds* λ_S as one moves across the family of threshold operators from the soft to the hard threshold operator. Figure 4 illustrates these phenomena for the MC+ penalty, before and after calibration.

Why the need for calibration? Because of the possibility of multiple stationary points in a non-convex optimization problem, good warm-starts are essential for avoiding sub-optimal solutions. The calibration assists in efficiently computing the

doubly-regularized (γ, λ) paths for the coefficient profiles. For fixed λ we start with the exact LASSO solution (large γ). This provides an excellent warm start for the problem with a slightly decreased value for γ. This is continued in a gradual fashion as γ approaches the best-subset problem. The *df* calibration provides a natural path across neighboring solutions in the (λ, γ) space:

- λ_S increases slowly, decreasing the size of the active set;
- at the same time, the active coefficients are shrunk successively less.

We find that the calibration keeps the algorithm away from sub-optimal stationary points and accelerates the speed of convergence.

4.3 Desirable Properties for a Family of Threshold Operators

Consider the family of threshold operators

$$\mathrm{S}_\gamma(\cdot, \lambda) : \Re \to \Re \quad \gamma \in (\gamma_0, \gamma_1).$$

Based on our observations on the properties and irregularities of the different penalties and their associated threshold operators, we have compiled a list of properties we consider desirable:

1. $\gamma \in (\gamma_0, \gamma_1)$ should *bridge* the gap between soft and hard thresholding, with the following continuity at the end points

$$\mathrm{S}_{\gamma_1}(\tilde{\beta}, \lambda) = \mathrm{sgn}(\tilde{\beta})(\tilde{\beta} - \lambda)_+ \text{ and } \mathrm{S}_{\gamma_0}(\tilde{\beta}; \lambda) = \tilde{\beta}\mathbf{I}(|\tilde{\beta}| \geq \lambda_H) \tag{15}$$

   where λ and λ_H correspond to the *shrinkage thresholds* of the soft and hard threshold operators respectively.

2. λ should control the effective *df* in the family $\mathrm{S}_\gamma(\cdot, \lambda)$; for fixed λ, the effective *df* of $\mathrm{S}_\gamma(\cdot, \lambda)$ for all values of γ should be the same.

3. For every fixed λ, there should be a strict nesting (increase) of the *shrinkage thresholds* λ_S as γ decreases from γ_1 to γ_0.

4. The map $\tilde{\beta} \mapsto \mathrm{S}_\gamma(\tilde{\beta}, \lambda)$ should be continuous.

5. The univariate penalized least squares function $Q^{(1)}(\beta)$ (7) should be convex for every $\tilde{\beta}$. This ensures that coordinate-wise procedures converge to a stationary point. In addition this implies continuity of $\tilde{\beta} \mapsto \mathrm{S}_\gamma(\tilde{\beta}, \lambda)$ in the previous item.

6. The function $\gamma \mapsto \mathrm{S}_\gamma(\cdot, \lambda)$ should be continuous on $\gamma \in (\gamma_0, \gamma_1)$. This assures a smooth transition as one moves across the family of penalized regressions, in constructing the family of regularization paths.

We believe these properties are necessary for a meaningful analysis for any generic non-convex penalty. Enforcing them will require re-parametrization and some restrictions in the family of penalties (and threshold operators) considered. In terms of the four families discussed in Section 2:

- The threshold operator induced by the SCAD penalty does not encompass the entire continuum from the soft to hard thresholding operators (all the others do).

- None of the penalties satisfy the nesting property of the shrinkage thresholds (item 3) or the degree of freedom calibration (item 2). In Section 5.1 we recalibrate the MC+ penalty so as to achieve these.

- The threshold operators induced by ℓ_γ and Log-penalties are not continuous (item 4); they are for MC+ and SCAD.

- $Q^{(1)}(\beta)$ is strictly convex for both the SCAD and the MC+ penalty. $Q^{(1)}(\beta)$ is non-convex for *every* choice of $(\lambda > 0, \gamma)$ for the ℓ_γ penalty and *some* choices of $(\lambda > 0, \gamma)$ for the log-penalty.

In this paper we explore the details of our approach through the study of the MC+ penalty; indeed, it is the only one of the four we consider for which all the properties above are achievable.

5 Illustration via the MC+ Penalty

Here we give details on the recalibration of the MC+ penalty, and the algorithm that results. The MC+ penalized univariate least-squares objective criterion is

$$Q^1(\beta) = \tfrac{1}{2}(\beta - \tilde{\beta})^2 + \lambda \int_0^{|\beta|} \left(1 - \frac{x}{\gamma\lambda}\right)_+ dx. \tag{16}$$

This can be shown to be convex for $\gamma \geq 1$, and non-convex for $\gamma < 1$. The minimizer for $\gamma > 1$ is a piecewise-linear thresholding function (see Figure 4, middle) given by

$$\mathrm{S}_\gamma(\tilde{\beta}, \lambda) = \begin{cases} 0 & \text{if } |\tilde{\beta}| \leq \lambda; \\ \mathrm{sgn}(\tilde{\beta}) \left(\frac{(|\tilde{\beta}| - \lambda}{1 - \frac{1}{\gamma}}\right) & \text{if } \lambda < |\tilde{\beta}| \leq \lambda\gamma; \\ \tilde{\beta} & \text{if } |\tilde{\beta}| > \lambda\gamma. \end{cases} \tag{17}$$

Observe that in (17), for fixed $\lambda > 0$,

$$\begin{aligned} &\text{as } \gamma \to 1+, \qquad \mathrm{S}_\gamma(\tilde{\beta}, \lambda) \to \mathrm{H}(\tilde{\beta}, \lambda), \\ &\text{as } \gamma \to \infty, \qquad \mathrm{S}_\gamma(\tilde{\beta}, \lambda) \to \mathrm{S}(\tilde{\beta}, \lambda). \end{aligned} \tag{18}$$

Hence $\gamma_0 = 1+$ and $\gamma_1 = \infty$. It is interesting to note here that the hard-threshold operator, which is conventionally understood to arise from a highly non-convex ℓ_0 penalized criterion (9), can be equivalently obtained as the limit of a sequence $\{\mathrm{S}_\gamma(\tilde{\beta}, \lambda)\}_{\gamma>1}$

where each threshold operator is the solution of a *convex criterion* (16) for $\gamma > 1$. For a fixed λ, this gives a family $\{\mathrm{S}_\gamma(\tilde{\beta}, \lambda)\}_\gamma$ with the soft and hard threshold operators as its two extremes.

5.1 Calibrating MC+ for *df*

For a fixed λ, we would like all the thresholding functions $\mathrm{S}_\gamma(\tilde{\beta}, \lambda)$ to have the same *df*; this will require a reparametrization. It turns out that this is tractable for the MC+ threshold function.

Consider the following univariate regression model

$$y_i = x_i\beta + \epsilon_i,\ i = 1, \ldots, n,\ \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2). \tag{19}$$

Without loss of generality, assume the x_i's have sample mean zero and sample variance one (the x_i's are assumed to be non-random). The *df* for the threshold operator $\mathrm{S}_\gamma(\cdot, \lambda)$ is defined by (14) with

$$\hat{\boldsymbol{\mu}}_{\gamma,\lambda} = \mathbf{x} \cdot \mathrm{S}_\gamma(\tilde{\beta}, \lambda) \tag{20}$$

and $\tilde{\beta} = \sum_i x_i y_i / \sum x_i^2$. The following theorem gives an explicit expression of the *df*.

Theorem 1. *For the model described in (19), the* df *of* $\hat{\boldsymbol{\mu}}_{\gamma,\lambda}$ *(20) is given by:*

$$df(\hat{\boldsymbol{\mu}}_{\gamma,\lambda}) = \frac{\lambda\gamma}{\lambda\gamma - \lambda}\Pr(\lambda \le |\tilde{\beta}| < \lambda\gamma) + \Pr(|\tilde{\beta}| > \lambda\gamma) \tag{21}$$

where these probabilities are to be calculated under the law $\tilde{\beta} \sim N(\beta, \sigma^2/n)$

Proof. We make use of Stein's unbiased risk estimation (Stein 1981, SURE) result. It states that if $\hat{\boldsymbol{\mu}} : \Re^n \to \Re^n$ is an almost differentiable function of $\mathbf{y}$, and $\mathbf{y} \sim$

$N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, then there is a simplification to the $df(\hat{\boldsymbol{\mu}})$ formula (14)

$$\begin{aligned} df(\hat{\boldsymbol{\mu}}) &= \sum_{i=1}^{n} \mathrm{Cov}(\hat{\mu}_i, y_i)/\sigma^2 \\ &= E(\nabla \cdot \hat{\boldsymbol{\mu}}), \end{aligned} \tag{22}$$

where $\nabla \cdot \hat{\boldsymbol{\mu}} = \sum_{i=1}^{n} \partial \hat{\mu}_i / \partial y_i$. Proceeding along the lines of proof in Zou et al. (2007), the function $\hat{\boldsymbol{\mu}}_{\gamma,\lambda} : \Re \to \Re$ defined in (20) is uniformly Lipschitz (for every $\lambda \geq 0, \gamma > 1$), and hence almost everywhere differentiable. The result follows easily from the definition (17) and that of $\tilde{\beta}$, which leads to (21). □

As $\gamma \to \infty$, $\hat{\boldsymbol{\mu}}_{\gamma,\lambda} \to \hat{\boldsymbol{\mu}}_{\lambda}$, and from (21), we see that

$$\begin{aligned} df(\hat{\boldsymbol{\mu}}_{\gamma,\lambda}) &\longrightarrow \Pr(|\tilde{\beta}| > \lambda) \\ &= E\left(I(|\tilde{\beta}| > 0)\right) \end{aligned} \tag{23}$$

which corresponds to the expression obtained by Zou et al. (2007) for the *df* of the LASSO in the univariate case.

Corollary 1. *The* df *for the hard-thresholding function* $\mathrm{H}(\tilde{\beta}, \lambda)$ *is given by*

$$df(\hat{\boldsymbol{\mu}}_{1+,\lambda}) = \lambda \phi^*(\lambda) + \Pr(|\tilde{\beta}| > \lambda) \tag{24}$$

where ϕ^* *is taken to be the p.d.f. of the absolute value of a normal random variable with mean* β *and variance* σ^2/n.

For the hard threshold operator, Stein's simplified formula does not work, since the corresponding function $\boldsymbol{y} \mapsto \mathrm{H}(\tilde{\beta}, \lambda)$ is not almost differentiable. But observing from (21) that

$$df(\hat{\boldsymbol{\mu}}_{\gamma,\lambda}) \to \lambda \phi^*(\lambda) + \Pr(|\tilde{\beta}| > \lambda), \text{ and } \mathrm{S}_{\gamma}(\tilde{\beta}, \lambda) \to \mathrm{H}(\tilde{\beta}, \lambda) \text{ as } \gamma \to 1+ \tag{25}$$

we get an expression for *df* as stated.

These expressions are consistent with simulation studies based on Monte Carlo estimates of *df*. Figure 4[Left] shows *df* as a function of γ for a fixed value $\lambda = 1$ for the uncalibrated MC+ threshold operators $\mathrm{S}_\gamma(\cdot, \lambda)$. For the figure we used $\beta = 0$ and $\sigma^2 = 1$.

5.1.1 Re-parametrization of the MC+ Penalty

We argued in Section 4.2 that for the *df* to remain constant for a fixed λ and varying γ, the shrinkage threshold $\lambda_S = \lambda_S(\lambda, \gamma)$ should increase as γ moves from γ_1 to γ_0. Theorem 3 formalizes this observation.

Hence for the purpose of *df* calibration, we re-parametrize the family of penalties as follows:

$$\lambda P^*(|t|; \lambda; \gamma) = \lambda_S(\lambda, \gamma) \int_0^{|t|} \left(1 - \frac{x}{\gamma \lambda_S(\lambda, \gamma)}\right)_+ dx. \tag{26}$$

With the re-parametrized penalty, the thresholding function is the obvious modification of (17):

$$\begin{aligned} \mathrm{S}^*_\gamma(\tilde{\beta}, \lambda) &= \arg\min_\beta \{\tfrac{1}{2}(\beta - \tilde{\beta})^2 + \lambda P^*(|\beta|; \lambda; \gamma)\} \\ &= \mathrm{S}_\gamma(\tilde{\beta}, \lambda_S(\lambda, \gamma)) \end{aligned} \tag{27}$$

Similarly for the *df* we simply plug into the formula (21)

$$df(\hat{\boldsymbol{\mu}}^*_{\gamma,\lambda}) = \frac{\gamma \lambda_S(\lambda, \gamma)}{\gamma \lambda_S(\lambda, \gamma) - \lambda_S(\lambda, \gamma)} \Pr(\lambda_S(\lambda, \gamma) \le |\tilde{\beta}| < \gamma \lambda_S(\lambda, \gamma)) + \Pr(|\tilde{\beta}| > \gamma \lambda_S(\lambda, \gamma)) \tag{28}$$

where $\hat{\boldsymbol{\mu}}^*_{\gamma,\lambda} = \hat{\boldsymbol{\mu}}_{\gamma, \lambda_S(\lambda,\gamma)}$.

To achieve a constant *df*, we require the following to hold for $\lambda > 0$:

- The shrinkage-threshold for the soft threshold operator is $\lambda_S(\lambda, \gamma = \infty) = \lambda$ and hence $\hat{\boldsymbol{\mu}}^*_\lambda = \hat{\boldsymbol{\mu}}_\lambda$ (a boundary condition in the calibration).
- $df(\hat{\boldsymbol{\mu}}^*_{\gamma,\lambda}) = df(\hat{\boldsymbol{\mu}}_\lambda)\ \forall\ \gamma > 1$.

The definitions for *df* depend on β and σ^2/n. Since the notion of *df* centers around variance, we use the null model with $\beta = 0$. We also assume without loss of generality that $\sigma^2/n = 1$, since this gets absorbed into λ.

Theorem 2. *For the calibrated MC+ penalty to achieve constant* df*, the shrinkage-threshold $\lambda_S = \lambda_S(\lambda, \gamma)$ must satisfy the following functional relationship*

$$\Phi(\gamma\lambda_S) - \gamma\Phi(\lambda_S) = -(\gamma - 1)\Phi(\lambda), \tag{29}$$

where Φ is the standard normal cdf and ϕ the pdf of the same.

Theorem 2 is proved in the Supplementary Materials Section 1.2, using (28) and the boundary constraint. The next theorem establishes some properties of of the map $(\lambda, \gamma) \mapsto (\lambda_S(\lambda, \gamma), \gamma\lambda_S(\lambda, \gamma))$, including the important nesting of shrinkage thresholds.

Theorem 3. *For a fixed λ,*

(a) $\gamma\lambda_S(\lambda, \gamma)$ is increasing as a function of γ.

(b) $\lambda_S(\lambda, \gamma)$ is decreasing as a function of γ

Note that as γ increases, we approach the LASSO; see Figure 4 (right). Both these relationships can be derived from the functional equation (29). Theorem 3 is proved in the Supplementary Materials Section 1.2.

5.1.2 Efficient Computation of *shrinkage thresholds*

In order to implement the calibration for MC+, we need an efficient method for evaluating $\lambda_S(\lambda, \gamma)$—i.e. solving equation (29). For this purpose we propose a simple parametric form for λ_S based on some of its required properties: the monotonicity properties just described, and the *df* calibration. We simply give the expressions

here, and leave the details for the Supplementary Materials Section 1.3:

$$\lambda_S(\lambda, \gamma) = \lambda_H(\frac{1-\alpha^*}{\gamma} + \alpha^*) \tag{30}$$

where

$$\alpha^* = \lambda_H^{-1}\Phi^{-1}(\Phi(\lambda_H) - \lambda_H\phi(\lambda_H)), \quad \lambda = \lambda_H\alpha^* \tag{31}$$

The above approximation turns out to be a reasonably good estimator for all practical purposes, achieving a calibration of *df* within an accuracy of five percent, uniformly over all (γ, λ). The approximation can be improved further, if we take this estimator as the starting point, and obtain recursive updates for $\lambda_S(\lambda, \gamma)$. Details of this approximation along with the algorithm are explained in the Supplementary Materials Section 1.3. Numerical studies show that a few iterations of this recursive computation can improve the degree of accuracy of calibration up to an error of 0.3 percent; Figure 4 [Left] was produced using this approximation. Figure 5 shows a typical pattern of recalibration for the example we present in Figure 7 in Section 8.

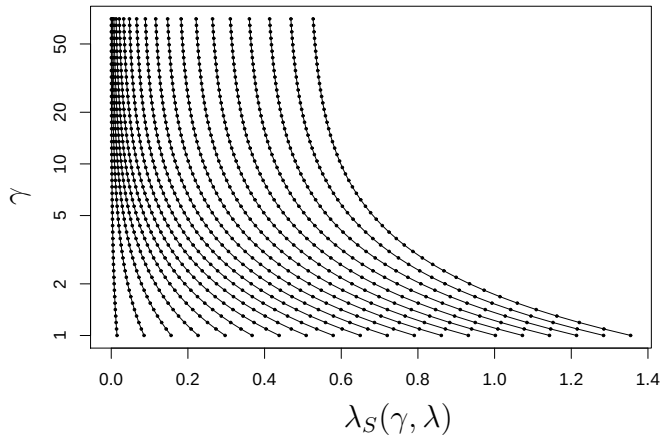

Figure 5: *Recalibrated values of λ via $\lambda_S(\gamma, \lambda)$ for the MC+ penalty. The values of λ at the top of the plot correspond to the LASSO. As γ decreases, the calibration increases the shrinkage threshold to achieve a constant univariate df.*

As noted below Algorithm 1 this leads to a lattice of values $\lambda_{k,\ell}$.

6 Convergence Analysis

In this section we present results on the convergence of Algorithm 1. Tseng & Yun (2009) show convergence of coordinate-descent for functions that can be written as the sum of a smooth function (loss) and a separable non-smooth convex function (penalty). This is not directly applicable to our case as the penalty is non-convex.

Denote the coordinate-wise updates $\boldsymbol{\beta}^{k+1} = \mathrm{S}_{\mathrm{cw}}(\boldsymbol{\beta}^k),\ k = 1, 2, \ldots$ with

$$\beta_j^{k+1} = \arg\min_u\ Q(\beta_1^{k+1}, \ldots, \beta_{j-1}^{k+1}, u, \beta_{j+1}^k, \ldots, \beta_p^k), j = 1, \ldots, p. \tag{32}$$

Theorem 4 establishes that under certain conditions, *SparseNet* always converges to a minimum of the objective function; conditions that are met, for example, by the SCAD and MC+ penalties (for suitable γ).

Theorem 4. *Consider the criterion in (6), where the given data $(y, \boldsymbol{X})$ lies on a compact set and no column of $\boldsymbol{X}$ is degenerate (ie multiple of the unit vector). Suppose the penalty $\lambda P(t; \lambda; \gamma) \equiv P(t)$ satisfies $P(t) = P(-t)$, $P'(|t|)$ is non-negative, uniformly bounded and $\inf_t P''(|t|) > -1$; where $P'(|t|)$ and $P''(|t|)$ are the first and second derivatives (assumed to exist) of $P(|t|)$ wrt $|t|$.*

Then the univariate maps $\beta \mapsto Q^{(1)}(\beta)$ are strictly convex and the sequence of coordinate-updates $\{\boldsymbol{\beta}^k\}_k$ converge to a minimum of the function $Q(\boldsymbol{\beta})$.

Note that the condition on the data $(y, \mathbf{X})$ is a mild assumption (as it is necessary for the variables to be standardized). Since the columns of $\mathbf{X}$ are mean-centered, the non-degeneracy assumption is equivalent to assuming that no column is identically zero.

The proof is provided in Appendix A.1. Lemma 1 and Lemma 2 under milder regularity conditions, establish that every limit point of the sequence $\{\boldsymbol{\beta}^k\}$ is a stationary point of $Q(\boldsymbol{\beta})$.

Remark 1. *Note that Theorem 4 includes the case where the penalty function $P(t)$ is convex in t.*

7 Simulation Studies

In this section we compare the simulation performance of a number of different methods with regard to a) prediction error, b) the number of non-zero coefficients in the model, and c) "misclassification error" of the variables retained. The methods we compare are *SparseNet*, Local Linear Approximation (LLA) and Multi-stage Local Linear Approximation (MLLA) (Zou & Li 2008, Zhang 2009, Candes et al. 2008), LASSO, forward-stepwise regression and best-subset selection. Note that *SparseNet*, LLA and MLLA are all optimizing the same MC+ penalized criterion.

We assume a linear model $Y = X\beta + \varepsilon$ with multivariate Gaussian predictors X and Gaussian errors. The Signal-to-Noise Ratio (SNR) and Standardized Prediction Error (SPE) are defined as

$$\text{SNR} = \frac{\sqrt{\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}}{\sigma}, \quad \text{SPE} = \frac{E(\mathbf{y} - \mathbf{x}\hat{\beta})^2}{\sigma^2} \tag{33}$$

The minimal achievable value for SPE is 1 (the Bayes error rate). For each model, the *optimal* tuning parameters — λ for the penalized regressions, subset size for best-subset selection and stepwise regression—are chosen based on minimization of the prediction error on a separate large validation set of size 10K. We use SNR=3 in all the examples.

Since a primary motivation for considering non-convex penalized regressions is to mimic the behavior of best subset selection, we compare its performance with best-subset for small $p = 30$. For notational convenience $\Sigma(\rho; m)$ denotes a $m \times m$ matrix with 1's on the diagonal, and ρ's on the off-diagonal.

We consider the following examples:

$\boldsymbol{S_1}$: $n = 35,\ p = 30,\ \Sigma^{S_1} = \Sigma(0.4; p)$ and $\boldsymbol{\beta}^{S_1} = (0.03, 0.07, 0.1, 0.9, 0.93, 0.97, \mathbf{0}_{1\times 24})$.

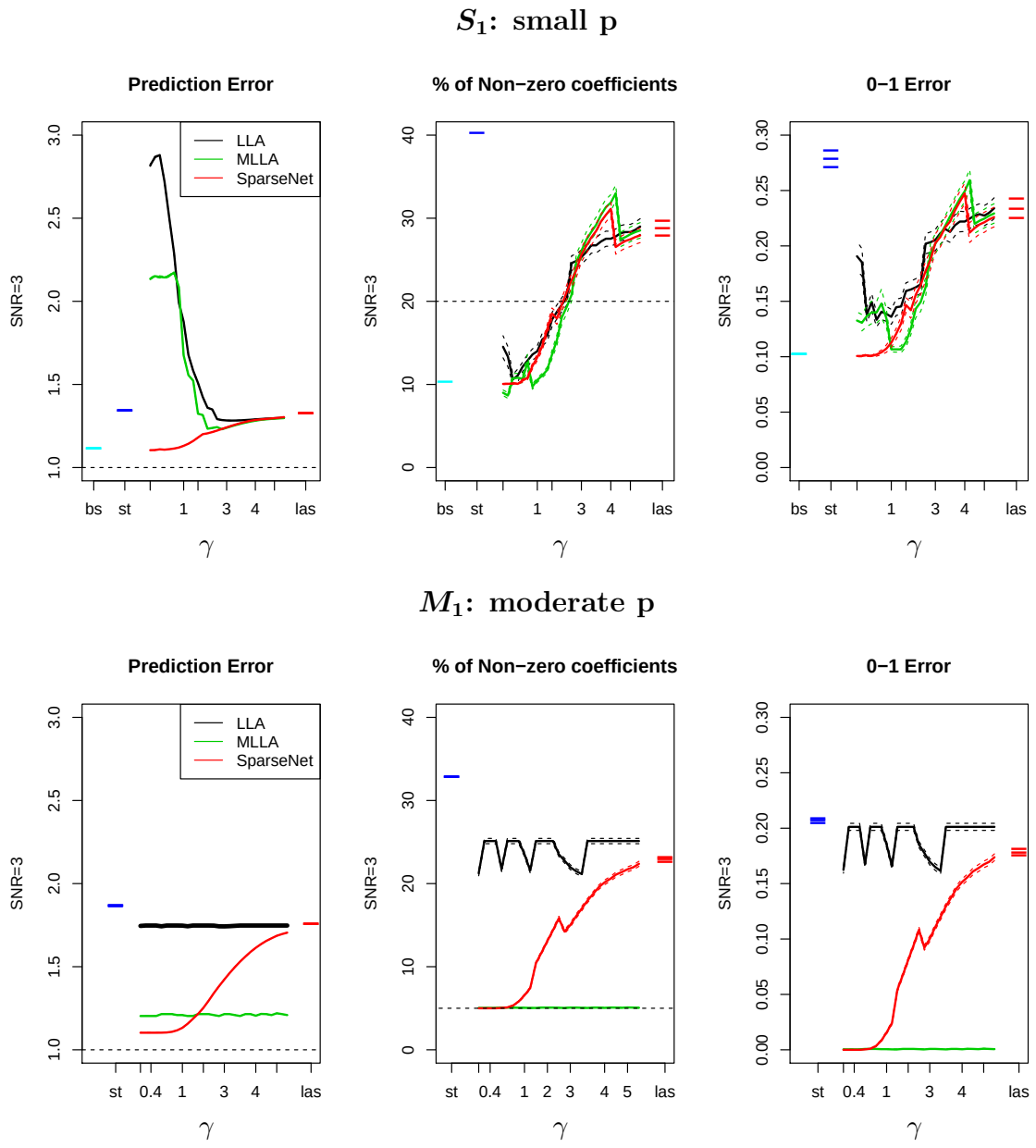

Figure 6: Examples $\boldsymbol{S_1}$ and $\boldsymbol{M_1}$. The three columns of plots represent prediction error, recovered sparsity and zero-one loss (all with standard errors over the 25 runs). Plots show the recalibrated MC+ family for different values of γ (on the log-scale) via *SparseNet*, LLA, MLLA, LASSO (las), step-wise (st) and best-subset (bs) (for $\boldsymbol{S_1}$ only)

$\boldsymbol{M_1}$: $n = 100,\ p = 200,\ \Sigma^{M_1} = \{0.7^{|i-j|}\}_{1\le i,j\le p}$ and $\boldsymbol{\beta}^{M_1}$ has 10 non-zeros such that $\beta^{M_1}_{20i+1} = 1,\ i = 0, 1, \ldots, 9$; and $\beta^{M_1}_i = 0$ otherwise.

In example $\boldsymbol{M_1}$ best-subset was not feasible and hence was omitted. The starting

point for both LLA and MLLA were taken as a vector of all ones (Zhang 2009). Results are shown in Figure 6 with discussion in Section 7.1. These two examples, chosen from a battery of similar examples, show that *SparseNet* compares very favorably with its competitors.

We now consider some larger examples and study the performance of *SparseNet* varying γ.

$\boldsymbol{M_1}(5)$: $n = 500,\ p = 1000,\ \Sigma^{M_1(5)} = \text{blockkdiag}(\Sigma^{M_1}, \ldots, \Sigma^{M_1})$ and $\boldsymbol{\beta}^{M_1(5)} = (\boldsymbol{\beta}^{M_1}, \ldots, \boldsymbol{\beta}^{M_1})$ (five blocks).

$\boldsymbol{M_1}(10)$: $n = 500,\ p = 2000$ (same as above with ten blocks instead of five).

$\boldsymbol{M_2}(5)$: $n = 500,\ p = 1000,\ \Sigma^{M_2(5)} = \text{blockdiag}(\Sigma(0.5, 200), \ldots, \Sigma(0.5, 200))$ and $\boldsymbol{\beta}^{M_2(5)} = (\boldsymbol{\beta}^{M_2}, \ldots, \boldsymbol{\beta}^{M_2})$ (five blocks). Here $\boldsymbol{\beta}^{M_2} = (\beta_1, \beta_2, \ldots, \beta_{10}, \mathbf{0}_{1\times 190})$ is such that the first ten coefficients form an equi-spaced grid on $[0, 0.5]$.

$\boldsymbol{M_2}(10)$: $n = 500,\ p = 2000$, and is like $\boldsymbol{M_2}(5)$ with ten blocks.

The results are summarized in Table 1, with discussion in Section 7.1.

7.1 Discussion of Simulation Results

In both $\boldsymbol{S_1}$ and $\boldsymbol{M_1}$, the aggressive non-convex *SparseNet* out-performs its less aggressive counterparts in terms of prediction error and variable selection. Due to the correlation among the features, the LASSO and the less aggressive penalties estimate a considerable proportion of zero coefficients as non-zero. *SparseNet* estimates for $\gamma \approx 1$ are almost identical to the best-subset procedure in $\boldsymbol{S_1}$. Step-wise performs worse in both cases. LLA and MLLA show similar behavior in $\boldsymbol{S_1}$, but are inferior to *SparseNet* in predictive performance for smaller values of γ. This shows that MLLA and *SparseNet* reach very different local optima. In $\boldsymbol{M_1}$ LLA/MLLA seem to show similar behavior across varying degrees of non-convexity. This is undesirable behavior, and is probably because MLLA/LLA gets stuck in a local minima. MLLA does perfect variable selection and shows good predictive accuracy.

Example: $\boldsymbol{M_1}(5)$	$\log(\gamma)$ values 3.92	1.73	1.19	0.10	Average std error
SPE	1.6344	1.3194	1.2015	1.1313	5.648×10^{-4}
% of non-zeros [5]	16.3	13.3	8.2	5.000	0.0702
$0-1$ error	0.1137	0.0825	0.0319	0.0002	8.508×10^{-4}
Example: $\boldsymbol{M_2}(5)$					
SPE	1.3797	1.467	1.499	1.7118	2.427×10^{-3}
% of non-zeros [5]	21.4	9.9	6.7	2.5	0.1037
$0-1$ error	0.18445	0.08810	0.0592	0.03530	1.118×10^{-3}
Example: $\boldsymbol{M_1}(10)$					
SPE	3.6819	4.4838	4.653	5.4729	1.152×10^{-2}
% of non-zeros [5]	23.0	12.0	8.5	6.2625	0.066
$0-1$ error	0.1873	0.0870	0.0577	0.0446	1.2372×10^{-3}
Example: $\boldsymbol{M_2}(10)$					
SPE	1.693	1.893	2.318	2.631	1.171×10^{-2}
% of non-zeros [5]	15.5	8.3	5.7	2.2	0.0685
$0-1$ error	0.1375	0.0870	0.0678	0.047	0.881×10^{-3}

Table 1: Table showing standardized prediction error (SPE), percentage of non-zeros and zero-one error in the recovered model via *SparseNet*, for different problem instances. The last column shows the averaged standard errors, across the four γ values. The true number of non-zero coefficients in the model are in square braces. Results are averaged over 25 runs.

In $\boldsymbol{M_1}(5)$, $(n = 500, p = 1000)$ the prediction error decreases steadily with decreasing γ, the variable selection properties improve as well. In $\boldsymbol{M_1}(10)$ $(n = 500, p = 2000)$ the prediction error increases overall, and there is a trend reversal — with the more aggressive non-convex penalties performing worse. The variable selection properties of the aggressive penalties, however, are superior to its counterparts. The less aggressive non-convex selectors include a larger number of variables with high shrinkage, and hence perform well in predictive accuracy.

In $\boldsymbol{M_2}(5)$ and $\boldsymbol{M_2}(10)$, the prediction accuracy decreases marginally with increasing γ. However, as before, the variable selection properties of the more aggressive non-convex penalties are far better.

In summary, *SparseNet* is able to mimic the prediction performance of best subset regression in these examples. In the high-p settings, it repeats this behavior, mimicking the best, and is the out-right winner in several situations since best-subset

regression is not available. In situations where the less aggressive penalties do well, *SparseNet* often reduces the number of variables for comparable prediction performance, by picking a γ in the interior of its range. The solutions of LLA and MLLA are quite different from the *SparseNet* and they often show similar performances across γ (as also pointed out in Candes et al. (2008) and Zou & Li (2008)).

It appears that the main differences among the different strategies MLLA, LLA and *SparseNet* lie in their roles of optimizing the objective $Q(\boldsymbol{\beta})$. We study this from a well grounded theoretical framework in Section 8, simulation experiments in Section 8.1 and further examples in Section 1.4 in the Supplementary Materials Section.

8 Other Methods of Optimization

We shall briefly review some of the state-of-the art methods proposed for optimization with general non-convex penalties.

Fan & Li (2001) used a local quadratic approximation (LQA) of the SCAD penalty. The method can be viewed as a majorize-minimize algorithm which repeatedly performs a weighted ridge regression. LQA gets rid of the non-singularity of the penalty at zero, depends heavily on the initial choice $\boldsymbol{\beta}^{\mathbf{0}}$, and is hence suboptimal (Zou & Li 2008) in searching for sparsity.

The MC+ algorithm of Zhang (2010) cleverly tracks multiple local minima of the objective function and seeks a solution to attain desirable statistical properties. The algorithm is complex, and since it uses a LARS-type update, is potentially slower than coordinate-wise procedures.

Friedman (2008) proposes a path-seeking algorithm for general non-convex penalties. Except in simple cases, it is unclear what criterion of the form "loss+penalty" it optimizes.

Multi-stage Local Linear Approximation. Zou & Li (2008), Zhang (2009) and

Candes et al. (2008) propose majorize-minimize (MM) algorithms for minimization of $Q(\boldsymbol{\beta})$ (6), and consider the re-weighted ℓ_1 criterion

$$Q(\boldsymbol{\beta};\boldsymbol{\beta}^k) = \text{Constant} + \tfrac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\sum_{j=1}^{p} P'(|\hat{\beta}_j^k|;\lambda;\gamma)|\beta_j| \tag{34}$$

for some starting point $\boldsymbol{\beta}^1$. If $\boldsymbol{\beta}^{k+1}$ is obtained by minimizing (34)— we denote the update via the map $\boldsymbol{\beta}^{k+1} := M_{\text{lla}}(\boldsymbol{\beta}^k)$. There is currently a lot of attention payed to MLLA in the literature, so we analyze and compare it to our coordinate-wise procedure in more detail.

We have observed (Sections 8.1 and 1.4.1) that the starting point is critical; see also Candes et al. (2008) and Zhang (2010). Zou & Li (2008) suggest stopping at $k = 2$ after starting with an "optimal" estimator, for example, the least-squares estimator when $p < n$. Choosing an optimal initial estimator $\boldsymbol{\beta}^1$ is essentially equivalent to knowing a-priori the zero and non-zero coefficients of $\boldsymbol{\beta}$. The adaptive LASSO of Zou (2006) is very similar in spirit to this formulation, though the weights (depending upon a good initial estimator) are allowed to be more general than derivatives of the penalty (34).

Convergence properties of *SparseNet* and MLLA can be analyzed via properties of fixed points of the maps $\mathrm{S}_{\text{cw}}(\cdot)$ and $M_{\text{lla}}(\cdot)$ respectively. If the sequences produced by $M_{\text{lla}}(\cdot)$ or $\mathrm{S}_{\text{cw}}(\cdot)$ converge, they correspond to stationary points of the function Q. Fixed points of the maps $M_{\text{lla}}(\cdot)$ and $\mathrm{S}_{\text{cw}}(\cdot)$ correspond to convergence of the sequence of updates.

The convergence analysis of MLLA is based on the fact that $Q(\boldsymbol{\beta};\boldsymbol{\beta}^k)$ majorizes $Q(\boldsymbol{\beta})$. If $\boldsymbol{\beta}^k$ is the sequence produced via MLLA, then $\{Q(\boldsymbol{\beta}^k)\}_{k\geq 1}$ converges. This does not address the convergence properties of $\{\boldsymbol{\beta}^k\}_{k\geq 1}$ nor properties of its limit points. Zou & Li (2008) point out that if the map M_{lla} satisfies $Q(\boldsymbol{\beta}) = Q(M_{\text{lla}}(\boldsymbol{\beta}))$ for limit points of the sequence $\{\boldsymbol{\beta}^k\}_{k\geq 1}$, then the limit points are stationary points of the function $Q(\boldsymbol{\beta})$. It appears, however, that solutions of $Q(\boldsymbol{\beta}) = Q(M_{\text{lla}}(\boldsymbol{\beta}))$ are not

easy to characterize by explicit regularity conditions on the penalty P. The analysis does not address the fate of all the limit points of the sequence $\{\boldsymbol{\beta}^k\}_{k\geq 1}$ in case the sufficient conditions of Proposition 1 in Zou & Li (2008) fail to hold true. Our analysis in Section 6 for *SparseNet* addresses all the concerns raised above. Fixed points of the map $\mathrm{S}_{\mathrm{cw}}(\cdot)$ are explicitly characterized through some regularity conditions of the penalty functions as described in Section 6. Explicit convergence can be shown under additional conditions — Theorem 4.

Furthermore, a fixed point of the map $M_{\mathrm{lla}}(\cdot)$ need not be a fixed point of $\mathrm{S}_{\mathrm{cw}}(\cdot)$ – for specific examples see Section 1.4 in the Supplementary Materials Section. Every stationary point of the coordinate-wise procedure, however, is also a fixed point of MLLA.

The formal framework being laid, we proceed to perform some simple numerical experiments in Section 8.1. Further discussions and comparisons between coordinate-wise procedures and MLLA can be found in Section 1.4 in the Supplementary Materials Section.

8.1 Empirical Performance and the Role of Calibration

In this section we show the importance of calibration and the specific manner in which the regularization path is traced out via *SparseNet* in terms of the quality of the solution obtained. In addition we compare the optimization performance of LLA/MLLA versus *SparseNet*.

In Figure 7 we consider an example with $n = p = 10$, $\Sigma = \{0.8^{|i-j|}\}$, $1 \leq i, j, \leq p$, SNR=1, $\beta_1 = \beta_p = 1$, all other $\beta_j = 0$. We consider 50 values of γ and 20 values of λ, and compute the solutions for all methods at the recalibrated pairs $(\gamma, \lambda_S(\gamma, \lambda))$; see Figure 5 for the exact $\lambda_{k,\ell}$ values used in these plots.

We compare *SparseNet* with

Type (a) Coordinate-wise procedure that computes solutions on a grid of λ for every fixed γ, with warm-starts across λ (from larger to smaller values). No *df*

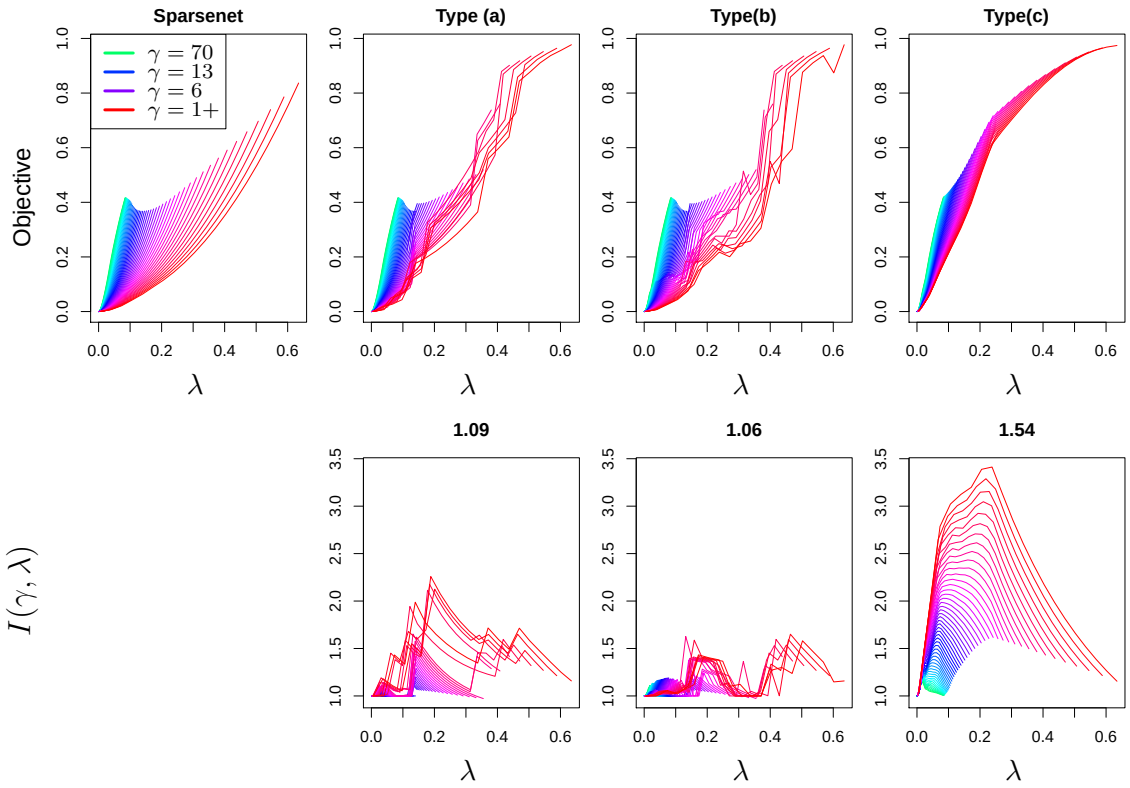

Figure 7: Top row: objective function for *SparseNet* compared to other coordinate-wise variants — Type (a), (b) and MLLA Type (c) — for a typical example. Plots are shown for 50 values of γ (some are labeled in the legend) and at each value of γ, 20 values of λ. Bottom row: relative increase $I(\gamma, \lambda)$ in the objective compared to *SparseNet*, with the average $\bar{I}$ reported at the top of each plot.

calibration is used here.

Type (b) Coordinate-wise procedure "cold-starting" with a zero vector (She 2009) for every pair (λ, γ).

Type (c) MLLA with the starting vector initialized to a vector of ones (Zhang 2009).

In all cases *SparseNet* shows superior performance. For each (γ, λ) pair we compute the relative increase $I(\gamma, \lambda) = [Q(\boldsymbol{\beta}^*) + 0.05]/[Q(\boldsymbol{\beta}^s) + 0.05]$, where $Q(\boldsymbol{\beta}^*)$ denotes the objective value obtained for procedure $*$ (types (a), (b) or (c) above), and $Q(\boldsymbol{\beta}^s)$ represents *SparseNet*. These values are shown in the second row of plots. They are also averaged over all values of (γ, λ) to give $\bar{I}$ (shown at the top of each plot). This is one of 24 different examples, over 6 different problem set-ups with 4 replicates in each. Figure 8 summarizes $\bar{I}$ for these 24 scenarios. Results are shown in Figure 7

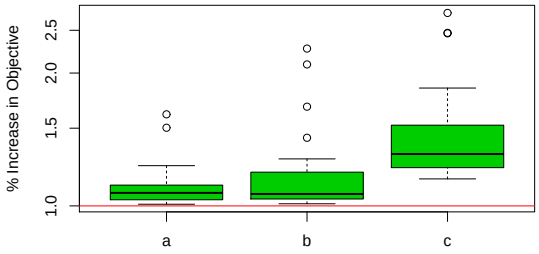

Figure 8: *Boxplots of $\bar{I}$ over 24 different examples. In all cases* SparseNet *shows superior performance.*

(We show the curves of the objective values, as in Figure 7, for all 24 scenarios in the Supplementary Materials Section.) The figure shows clearly the benefits of using the proposed calibration in terms of optimization performance. MLLA appears to get "stuck" in local optima, supporting our claim in section 7.1. For larger values of γ the different coordinate-wise procedures perform similarly, since the problem is close to convex.

A Appendix

A.1 Convergence Analysis for Algorithm 1

To prove Theorem 4 we require Lemma 1.

Lemma 1. *Suppose the data (y, X) lies on a compact set. In addition, we assume the following*

1. *The penalty function $P(t)$ (symmetric around 0) is differentiable on $t \geq 0$; $P'(|t|)$ is non-negative, continuous and uniformly bounded, where $P'(|t|)$ is the derivative of $P(|t|)$ wrt $|t|$.*

2. *The sequence generated $\{\boldsymbol{\beta}^k\}_k$ is bounded*

3. *For every convergent subsequence $\{\boldsymbol{\beta}^{n_k}\}_k \subset \{\boldsymbol{\beta}^k\}_k$ the successive differences converge to zero: $\boldsymbol{\beta}^{n_k} - \boldsymbol{\beta}^{n_k-1} \to 0$*

If $\boldsymbol{\beta}^{\infty}$ is any limit point of the sequence $\{\boldsymbol{\beta}^k\}_k$, then $\boldsymbol{\beta}^{\infty}$ is a minimum for the function $Q(\boldsymbol{\beta})$; i.e. for any $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p) \in \Re^p$

$$\lim_{\alpha \downarrow 0+} \inf \left\{ \frac{Q(\boldsymbol{\beta}^{\infty} + \alpha\boldsymbol{\delta}) - Q(\boldsymbol{\beta}^{\infty})}{\alpha} \right\} \geq 0 \tag{35}$$

In the Supplementary Materials Section 1 we present the proof of Lemma 1. We also show via Lemma 1 that assumption 2 is by no means restrictive. Lemma 2 gives a sufficient condition under which assumption 3 of Lemma 1 is true.

It is useful to note here that assumptions 2, 3 of Lemma 1 follow from the assumptions of Theorem 4 (we make a further note of this in the proof below).

We are now in a position to present the proof of Theorem 4:

Proof. **Proof of Theorem 4** We will assume WLOG that the data is standardized.

For fixed i and $(\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p)$, we will write $\chi^i_{(\beta_1,\ldots,\beta_{i-1},\beta_{i+1},\ldots,\beta_p)} \equiv \chi(u)$ for the sake notational convenience.

The sub-gradient(s) (Borwein & Lewis 2006) of $\chi(u)$ at u are given by

$$\partial\chi(u) = \nabla_i f((\beta_1, \ldots, \beta_{i-1}, u, \beta_{i+1}, \ldots, \beta_p)) + P'(|u|)\,\mathrm{sgn}(u) \tag{36}$$

Also observe that

$$\begin{aligned}
\chi(u+\delta) - \chi(u) &= \{f((\beta_1, \ldots, \beta_{i-1}, u+\delta, \beta_{i+1}, \ldots, \beta_p)) - \\
&\quad f((\beta_1, \ldots, \beta_{i-1}, u, \beta_{i+1}, \ldots, \beta_p))\} + \{P(|u+\delta|) - P(|u|)\} \quad (37) \\
&= \{\nabla_i f((\beta_1, \ldots, \beta_{i-1}, u, \beta_{i+1}, \ldots, \beta_p))\delta + \tfrac{1}{2}\nabla_i^2 f \delta^2\} \\
&+ \{P'(|u|)(|u+\delta| - |u|) + \tfrac{1}{2}P''(|u^*|)(|u+\delta| - |u|)^2\} \quad (38)
\end{aligned}$$

Line (38) is obtained from (37) by a Taylor's series expansions on f (wrt to the i^{th} coordinate) and $|t| \mapsto P(|t|)$. $\nabla_i^2 f$ is the second derivative of the function f wrt the i^{th} coordinate and equals 1 as the features are all standardized; $|u^*|$ is some number between $|u+\delta|$ and

$|u|$. The rhs of (38) can be simplified as follows:

$$\begin{aligned} & \nabla_i f((\beta_1,\ldots,\beta_{i-1},u,\beta_{i+1},\ldots,\beta_p))\delta + \tfrac{1}{2}\nabla_i^2 f\delta^2 \\ & + P'(|u|)(|u+\delta|-|u|) + \tfrac{1}{2}P''(|u^*|)(|u+\delta|-|u|)^2 \\ = \;& \tfrac{1}{2}\nabla_i^2 f\delta^2 + \{\nabla_i f((\beta_1,\ldots,\beta_{i-1},u,\beta_{i+1},\ldots,\beta_p))\delta + P'(|u|)\operatorname{sgn}(u)\delta\} + \\ & \{P'(|u|)(|u+\delta|-|u|) - P'(|u|)\operatorname{sgn}(u)\delta\} + \tfrac{1}{2}P''(|u^*|)(|u+\delta|-|u|)^2 \end{aligned} \tag{39}$$

Since $\chi(u)$ is minimized at u_0; $0 \in \partial\chi(u_0)$. Thus using (36) we get

$$\nabla_i f((\beta_1,\ldots,\beta_{i-1},u_0,\beta_{i+1},\ldots,\beta_p))\delta + P'(|u_0|)\operatorname{sgn}(u_0)\delta = 0 \tag{40}$$

if $u_0 = 0$ then the above holds true for some value of $\operatorname{sgn}(u_0) \in [-1,1]$.

By definition of sub-gradient of $x \mapsto |x|$ and $P'(|x|) \geq 0$ we have

$$P'(|u|)(|u+\delta|-|u|) - P'(|u|)\operatorname{sgn}(u)\delta = P'(|u|)\left((|u+\delta|-|u|) - \operatorname{sgn}(u)\delta\right) \geq 0 \tag{41}$$

Using (40,41) in (39,38) at $u = u_0$ we have

$$\chi(u_0+\delta) - \chi(u_0) \geq \tfrac{1}{2}\nabla_i^2 f\delta^2 + \tfrac{1}{2}P''(|u^*|)(|u_0+\delta|-|u_0|)^2 \tag{42}$$

Observe that $(|u_0+\delta|-|u_0|)^2 \leq \delta^2$. If $P''(|u^*|) \leq 0$ then $\frac{1}{2}P''(|u^*|)(|u_0+\delta|-|u_0|)^2 \geq \frac{1}{2}P''(|u^*|)\delta^2$. If $P''(|u^*|) \geq 0$ then $\frac{1}{2}P''(|u^*|)(|u_0+\delta|-|u_0|)^2 \geq 0$.

Combining these two we get

$$\chi(u_0+\delta) - \chi(u_0) \geq \tfrac{1}{2}\delta^2\left(\nabla_i^2 f + \min\{P''(|u^*|),0\}\right) \tag{43}$$

By the conditions of this theorem, $(\nabla_i^2 f + \inf_x P''(|x|)) > 0$. Hence there exists a $\theta > 0$,

$$\theta = \tfrac{1}{2}\{\nabla_i^2 f + \min\{\inf_x P''(|x|),0\}\} = \tfrac{1}{2}\{1 + \min\{\inf_x P''(|x|),0\}\} > 0$$

which is independent of $(\beta_1,\ldots,\beta_{i-1},\beta_{i+1},\ldots,\beta_p)$, such that

$$\chi(u_0+\delta) - \chi(u_0) \geq \theta\delta^2 \tag{44}$$

(For the MC+ penalty, $\inf_x P''(|x|) = -\frac{1}{\gamma};\ \gamma > 1$. For the SCAD, $\inf_x P''(|x|) = -\frac{1}{\gamma-1};\ \gamma > 2$.)

In fact the dependence on u_0 above in (44) can be further relaxed. Following the above arguments carefully, we see that the function $\chi(u)$ is strictly convex (Borwein & Lewis 2006) as it satisfies:

$$\chi(u+\delta) - \chi(u) - \partial\chi(u)\delta \geq \theta\delta^2 \tag{45}$$

We will use this result to show the convergence of $\{\boldsymbol{\beta}^k\}$. Using (44) we have

$$\begin{aligned} Q(\boldsymbol{\beta}_i^{m-1}) - Q(\boldsymbol{\beta}_{i+1}^{m-1}) &\geq \theta(\beta_{i+1}^{m-1} - \beta_{i+1}^{m})^2 \\ &= \theta\|\boldsymbol{\beta}_i^{m-1} - \boldsymbol{\beta}_{i+1}^{m-1}\|_2^2 \end{aligned} \tag{46}$$

where $\boldsymbol{\beta}_i^{m-1} := (\beta_1^m, \ldots, \beta_i^m, \beta_{i+1}^{m-1}, \ldots, \beta_p^{m-1})$. (46) shows the boundedness of the sequence $\boldsymbol{\beta}^m$, for every $m > 1$ since the starting point $\boldsymbol{\beta}^1 \in \Re^p$.

Using (46) repeatedly across every coordinate, we have for every m

$$Q(\boldsymbol{\beta}^{m+1}) - Q(\boldsymbol{\beta}^m) \geq \theta\|\boldsymbol{\beta}^{m+1} - \boldsymbol{\beta}^m\|_2^2 \tag{47}$$

Using the fact that the (decreasing) sequence $Q(\boldsymbol{\beta}^m)$ converges, it is easy to see from (47), that the sequence $\boldsymbol{\beta}^k$ cannot cycle without convergence ie it must have a unique limit point. This completes the proof of convergence of $\boldsymbol{\beta}^k$.

Observe that, (47) implies that assumptions 2 and 3 of Lemma 1 are satisfied. Hence using Lemma 1, the limit of $\boldsymbol{\beta}^k$ is a minimum of $Q(\boldsymbol{\beta})$ — this completes the proof. □

Supplementary Materials

Title: Supplementary Materials Some details of technical derivations and experiments are available in the Supplementary Materials Section. Please include the file

`jasa_MHF_suppl`

here.

References

Borwein, J. & Lewis, A. (2006), *Convex Analysis and Nonlinear Optimization*, Springer.

Breiman, L. (1996), 'Heuristics of instability and stabilization in model selection', *Annals of Statistics* **24**, 2350–2383.

Candes, E. J., Wakin, M. B. & Boyd, S. (2008), 'Enhancing sparsity by reweighted l1 minimization', *Journal of Fourier Analysis and Applications* **14**(5), 877–905.

Chen, S. & Donoho, D. (1994), On basis pursuit, Technical report, Department of Statistics Stanford University.

Donoho, D. (2006), 'For most large underdetermined systems of equations, the minimal ℓ^1-norm solution is the sparsest solution', *Communications on Pure and Applied Mathematics* **59**, 797–829.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression (with discussion)', *Annals of Statistics* **32**(2), 407–499.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360(13).

Frank, I. & Friedman, J. (1993), 'A statistical view of some chemometrics regression tools (with discussion)', *Technometrics* **35**(2), 109–148.

Friedman, J. (2008), Fast sparse regression and classification, Technical report, Department of Statistics, Stanford University.

Friedman, J., Hastie, T. & Tibshirani, R. (2009), *glmnet: Lasso and elastic-net regularized generalized linear models*. R package version 1.1-4.
URL: *http://www-stat.stanford.edu/ hastie/Papers/glmnet.pdf*

Gao, H.-Y. & Bruce, A. G. (1997), 'Waveshrink with firm shrinkage', *Statistica Sinica* **7**, 855–874.

Knight, K. & Fu, W. (2000), 'Asymptotics for lasso-type estimators', *Annals of Statistics* **28**(5), 1356–1378.

Meinshausen, N. & Bühlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *Annals of Statistics* **34**, 1436–1462.

Nikolova, M. (2000), 'Local strong homogeneity of a regularized estimator', *SIAM J. Appl. Math.* **61**, 633–658.

Osborne, M., Presnell, B. & Turlach, B. (2000), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**, 389–404.

She, Y. (2009), 'Thresholding-based iterative selection procedures for model selection and shrinkage', *Electronic Journal of Statistics* .

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *Annals of Statistics* **9**, 1131–1151.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Tseng, P. (2001), 'Convergence of a block coordinate descent method for nondifferentiable minimization', *Journal of Optimization Theory and Applications* **109**, 475–494.

Tseng, P. & Yun, S. (2009), 'A coordinate gradient descent method for nonsmooth separable minimization', *Mathematical Programming B* **117**, 387–423.

Zhang, C. H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *Annals of Statistics (to appear)* .

Zhang, C.-H. & Huang, J. (2008), 'The sparsity and bias of the lasso selection in high-dimensional linear regression', *Annals of Statistics* **36**(4), 1567–1594.

Zhang, T. (2009), Multi-stage convex relaxation for non-convex optimization, Technical report, Rutgers Tech Report.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2563.

Zou, H. (2006), 'The adaptive lasso and its oracle properties', *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H., Hastie, T. & Tibshirani, R. (2007), 'On the degrees of freedom of the lasso', *Annals of Statistics* **35**(5), 2173–2192.

Zou, H. & Li, R. (2008), 'One-step sparse estimates in nonconcave penalized likelihood models', *The Annals of Statistics* **36**(4), 1509–1533.