Electronic Journal of Statistics
Vol. 4 (2010) 1258–1299
ISSN: 1935-7524
DOI: 10.1214/10-EJS582

Majorization-Minimization algorithms for nonsmoothly penalized objective functions

Elizabeth D. Schifano, Robert L. Strawderman and Martin T. Wells

Department of Statistical Science
Cornell University, Ithaca NY 14853 USA
e-mail: eds27@cornell.edu; rls54@cornell.edu; mtw1@cornell.edu

Abstract: The use of penalization, or regularization, has become common in high-dimensional statistical analysis, where an increasingly frequent goal is to simultaneously select important variables and estimate their effects. It has been shown by several authors that these goals can be achieved by minimizing some parameter-dependent "goodness-of-fit" function (e.g., a negative loglikelihood) subject to a penalization that promotes sparsity. Penalty functions that are singular at the origin have received substantial attention, arguably beginning with the Lasso penalty [62].

The current literature tends to focus on specific combinations of differentiable goodness-of-fit functions and penalty functions singular at the origin. One result of this combined specificity has been a proliferation in the number of computational algorithms designed to solve fairly narrow classes of optimization problems involving objective functions that are not everywhere continuously differentiable. In this paper, we propose a general class of algorithms for optimizing an extensive variety of nonsmoothly penalized objective functions that satisfy certain regularity conditions. The proposed framework utilizes the majorization-minimization (MM) algorithm as its core optimization engine. In the case of penalized regression models, the resulting algorithms employ iterated soft-thresholding, implemented componentwise, allowing for fast and stable updating that avoids the need for inverting high-dimensional matrices. We establish convergence theory under weaker assumptions than previously considered in the statistical literature. We also demonstrate the exceptional effectiveness of new acceleration methods, originally proposed for the EM algorithm, in this class of problems. Simulation results and a microarray data example are provided to demonstrate the algorithm's capabilities and versatility.

AMS 2000 subject classifications: Primary 65C60, 62J07; secondary 62J05, 62J12.
Keywords and phrases: Convex optimization, iterative soft thresholding, Lasso penalty, minimax concave penalty, non-convex optimization, smoothly clipped absolute deviation penalty.

Received February 2010.

Contents

1. Introduction

Variable selection remains an important and challenging issue in statistics. Modern methods, increasingly based on the principle of penalized likelihood estimation and often used in high dimensional regression problems, attempt to achieve this goal through an adaptive variable selection process that simultaneously permits estimation of regression effects. The literature on penalized minimization of a "goodness-of-fit" function (e.g., negative loglikelihood), with a penalty singular at the origin, has become vast and continues to proliferate in part due to the consideration of specific combinations of goodness-of-fit and penalty functions, the associated statistical properties of resulting estimators, and the development of several combination-specific optimization algorithms, [e.g., 21, 24, 51, 62, 74, 75, 78].

Our primary goal in this paper is to propose a unified optimization framework that utilizes the Majorization-Minimization (MM) algorithm [e.g., 32, 38, 39] as the primary optimization tool. The resulting class of algorithms is referred to as MIST, an acronym for Minimization by Iterative Soft Thresholding. The MM algorithm has been considered previously in solving specific classes of singularly penalized likelihood estimation problems [e.g., 16, 33, 77]; to a large extent, this work is motivated by these ideas. Important advantages of the proposed work include the exceptional versatility of the class of MIST algorithms, their associated ease of implementation and numerical stability, and the availability of a fixed point convergence theory that permits weaker assumptions than existing papers in this area. We emphasize that the focus of this paper is on the development of a stable and versatile class of algorithms applicable to a wide variety of singularly penalized estimation problems. In particular, the consideration of asymptotic and oracle properties of estimators, as well as methods for effectively choosing associated penalty parameters, are not focal points of this paper. A

reasonably comprehensive treatment of these results may be found in Johnson, Lin and Zeng [34], where asymptotics and oracle properties for estimators derived from a general class of penalized estimating equations are developed in some detail.

The paper is organized as follows. In Section 2, we provide some general background on the class of MM algorithms. Section 2.2, in particular, introduces important notation and summarizes a set of useful sufficient conditions for local convergence of general MM algorithms applied to a large and interesting class of penalized optimization problems. In Section 3, we present a more specialized version of the general algorithm and show how the minimization step of the MM algorithm can be carried out using iterated soft-thresholding. In its most general form, iterated soft-thresholding is required at each minimization step. However, in the context of penalized estimation for the class of generalized linear regression models, we further show that a judicious choice of majorization function allows one to carry out this minimization step componentwise and in one iteration. Simulation results are provided in Section 4 and an application in survival analysis to Diffuse Large B Cell Lymphoma expression data [54] is presented in Section 5. We conclude with a discussion in Section 6. Proofs and other relevant results are collected in the Appendix.

2. Convergence of Majorization-Minimization algorithms

2.1. Review

A Majorization-Minimization (MM) algorithm is not really a single algorithm but rather a term that more aptly describes a general principle for solving a difficult minimization problem by transferring this problem to a related surrogate function that is much easier to minimize [39]. The acronym MM, as pointed out in the rejoinder to the discussion of Lange, Hunter and Yang [39], can also stand for "Minorization-Maximization" if one desires to maximize, rather than minimize, an objective function. The Expectation Maximization (EM) algorithm [18], originally developed in the context of missing data applications, is an important special case of the class of minorization-maximization algorithms [39]. As shown in Becker, Yang and Lange [6], the "optimization transfer" principle that underlies its construction exists independently of the missing data setting and leads to a powerful and general tool for constructing algorithms. Lange, Hunter and Yang [39] attribute one of the earliest examples of this class of algorithms to Ortega and Rheinboldt [50] as well as identify several later statistically-oriented examples of algorithms falling into this class.

From this point onward, we consider the "Majorization-Minimization" form of the MM algorithm. For simplicity of presentation, we develop all results for the problem of unconstrained minimization; our results can be extended to the problem of constrained minimization with minor changes to the proposed algorithms and more substantial changes to certain regularity assumptions and technical arguments. Let $\xi(\boldsymbol{\beta})$ denote a real-valued objective function to be

minimized for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ in $\mathbb{R}^p$. Let $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ denote some real-valued "surrogate" objective function, where $\boldsymbol{\alpha} \in \mathbb{R}^p$ is a bounded vector (i.e., a vector with bounded elements) with the same dimension as $\boldsymbol{\beta}$. Let

$$M(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \quad \xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha}); \tag{1}$$

then, if $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ majorizes $\xi(\boldsymbol{\beta})$ for each $\boldsymbol{\alpha}$, i.e.,

$$\xi(\boldsymbol{\beta}) = \xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\beta}) \text{ for each } \boldsymbol{\beta} \quad \text{and} \quad \xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \geq \xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\beta}) \text{ for } \boldsymbol{\beta} \neq \boldsymbol{\alpha},$$

a MM algorithm for minimizing $\xi(\boldsymbol{\beta})$ takes the following form:

1. Initialize the algorithm with $\boldsymbol{\beta}^{(0)}$, a bounded vector;
2. For $n \geq 0$, compute $\boldsymbol{\beta}^{(n+1)} = M(\boldsymbol{\beta}^{(n)})$, iterating until convergence.

Since $\boldsymbol{\beta}^{(n+1)}$ is the minimum of the surrogate function at $\boldsymbol{\beta}^{(n)}$, the MM procedure forces $\xi(\boldsymbol{\beta})$ downhill at each iteration, i.e., $\xi(\boldsymbol{\beta}^{(n+1)}) \leq \xi(\boldsymbol{\beta}^{(n)})$ for every $n \geq 0$.

Provided that the objective function, its surrogate and the mapping $M(\cdot)$ satisfy certain regularity conditions, one can also establish "convergence" of this algorithm. For example, Lange [38, Proposition 10.3.4] proves convergence of the MM iteration sequence assuming, among other things, that the objective functions $\xi(\boldsymbol{\beta})$ and $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are twice continuously differentiable; see Lange [37, Proposition 6] for related results on the generalized EM algorithm. However, weaker statements of convergence are also possible in problems that lack this degree of smoothness. For example, Lange, Hunter and Yang [39, Sec. 3] summarize conditions under which any limit point of the sequence $\boldsymbol{\beta}^{(n+1)} = M(\boldsymbol{\beta}^{(n)})$ is also a stationary point of some continuous objective function $\xi(\boldsymbol{\beta})$; see Lange [38, Propositions 10.3.1–10.3.3] and Lange [37, Propositions 3-5] for related theoretical developments. Convergence, at least as ordinarily interpreted, can be expected provided that $\xi(\boldsymbol{\beta})$ has a unique minimum; however, more generally, such "convergence" results do not necessarily imply that the MM iteration sequence itself converges to a unique limit.

This last observation is relevant to the convergence analysis of algorithms designed to solve singularly penalized regression problems. Common and increasingly important examples lacking the differentiability requirements of Lange [38, Proposition 10.3.4] include all penalized regression problems involving the lasso penalty [62], adaptive lasso penalty [74], elastic net penalty [75], and the smoothly clipped absolute deviation (SCAD) penalty [21]. In order to properly analyze algorithmic convergence in these settings, an appropriately general theory of convergence is required. One such theory is developed in Appendix A.1 and complements existing convergence theory for the EM and MM algorithms that may be found in Wu [68], Lange [37], Lange, Hunter and Yang [39], Tseng [64], Lange [38] and Chrétien and Hero [11], among other places. Two important contributions of these results include useful refinements of existing theory for the EM and MM algorithms as well as a set of sufficient conditions that are relatively straightforward to verify for the general class of penalized optimization problems considered in the next section.

2.2. Objective functions with nondifferentiable, separable penalties

In this section, we summarize convergence results for generic MM algorithms that are intended to minimize an objective function of the form

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

where $g(\boldsymbol{\beta})$ is a continuous "goodness of fit" function (e.g., a negative loglikelihood), $p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is a continuous but non-differentiable penalty function, and $\|\cdot\|$ denotes the usual Euclidean vector norm. Further regularity conditions will be given below; as shown later, the resulting class of problems represented by (2) contains the vast majority of penalized regression problems currently under investigation in the statistics literature. It also covers numerous additional problems by expanding the class of permissible goodness-of-fit and penalty functions in a substantial way.

We assume throughout that $g(\boldsymbol{\beta})$ is convex with at least one bounded local minimizer. This implies that $g(\boldsymbol{\beta})$ is coercive [e.g., 38, Chapter 10]; that is, $g(\boldsymbol{\beta})$ becomes unbounded as $\|\boldsymbol{\beta}\| \to \infty$. The convexity of $g(\boldsymbol{\beta})$ further implies that $g(\boldsymbol{\beta})$ is Lipschitz continuous on each compact subset of $\mathbb{R}^p$ (i.e., locally Lipschitz continuous). It follows that $\nabla g(\boldsymbol{\beta})$ exists for almost all $\boldsymbol{\beta}$. We further assume

$$p(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \sum_{j=1}^{p} \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j), \tag{3}$$

where the vector $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^T, \ldots, \boldsymbol{\lambda}_p^T)^T$ and $\boldsymbol{\lambda}_j$ denotes the block of $\boldsymbol{\lambda}$ associated with β_j. It is assumed that each $\boldsymbol{\lambda}_j$ has dimension greater than or equal to one, that all blocks have the same dimension, and that the $\boldsymbol{\lambda}_{j1} = \lambda$ for each $j \geq 1$. Evidently, the case where $\dim(\boldsymbol{\lambda}_j) = 1$ for $j \geq 1$ simply corresponds to the setting in which each coefficient is penalized in exactly the same way; permitting the dimension of $\boldsymbol{\lambda}_j$ to exceed one allows the penalty to depend on additional parameters (e.g., weights, such as in the case of the adaptive lasso considered in Zou [74]). We are interested in problems with penalization; therefore, λ is assumed bounded and strictly positive throughout this paper. Several specific examples will be discussed below. For a bounded vector $\boldsymbol{\theta}$ having $\lambda > 0$ as its first element, and the remainder of $\boldsymbol{\theta}$ collecting any additional parameters used to define the penalty, the scalar function $\tilde{p}(r; \boldsymbol{\theta})$ is assumed to satisfy the following condition:

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

Evidently, (P1) implies that $\tilde{p}'(r; \boldsymbol{\theta}) > 0$ for $r \in (0, K_{\boldsymbol{\theta}})$, where $K_{\boldsymbol{\theta}} > 0$ may be finite or infinite. The combination of the concavity and nonnegative derivative conditions imply that the penalty increases away from the origin, but with a decreasing rate of growth that may become zero. The case where (3) is identically zero for $r > 0$ is ruled out by the positivity of the right derivative at the origin;

similarly, the concavity assumption also rules out the possibility of a strictly convex penalty term. Neither of these restrictions is particularly problematic. Our specific interest lies in the development of algorithms for estimation problems subject to a penalty singular at the origin. Were (3) absent, or replaced by a strictly convex penalty term, the convexity of $g(\boldsymbol{\beta})$ implies (2) can be minimized directly using any suitable convex optimization algorithm, such as that discussed in Theorem 3.2 below.

Under the conditions specified above, the objective function (2) is not necessarily convex and may have multiple local minima. Theorem 2.1 establishes local convergence of the indicated class of MM algorithms for minimizing objective functions of the form (2). A proof is provided in Appendix A.2, where it is shown that the conditions imposed in the statement of the theorem are sufficient conditions for the application of the general MM local convergence theory summarized in Appendix A.1.

Theorem 2.1. *Let $g(\cdot)$ and $p(\cdot;\boldsymbol{\lambda})$ satisfy the indicated assumptions. Let $h(\boldsymbol{\beta},\boldsymbol{\alpha}) \geq 0$ be a real-valued, continuous function of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ that is continuously differentiable in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha}$ and satisfies $h(\boldsymbol{\beta},\boldsymbol{\alpha}) = 0$ when $\boldsymbol{\beta} = \boldsymbol{\alpha}$. Let*

$$q(\boldsymbol{\beta},\boldsymbol{\alpha};\boldsymbol{\lambda}) = \sum_{j=1}^{p} \tilde{q}(|\beta_j|,|\alpha_j|;\boldsymbol{\lambda}_j), \tag{4}$$

where $\tilde{q}(r,s;\boldsymbol{\theta}) = \tilde{p}(s;\boldsymbol{\theta}) + \tilde{p}'(s;\boldsymbol{\theta})(r-s)$ for $r,s \geq 0$. Assume $\mathcal{S}$, the set of stationary points for $\xi(\boldsymbol{\beta})$, is both non-empty and finite, where the notion of a stationary point is defined as in Clarke [12]. Then:

(i) $\xi(\boldsymbol{\beta})$ in (2) is locally Lipschitz continuous.
(ii) $q(\boldsymbol{\beta},\boldsymbol{\alpha};\boldsymbol{\lambda}) - p(\boldsymbol{\beta};\boldsymbol{\lambda}) \geq 0$ for all $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$ (possibly, identically zero).
(iii) $\xi^{[S]}(\boldsymbol{\beta},\boldsymbol{\alpha}) \equiv \xi(\boldsymbol{\beta}) + h(\boldsymbol{\beta},\boldsymbol{\alpha}) + q(\boldsymbol{\beta},\boldsymbol{\alpha};\boldsymbol{\lambda}) - p(\boldsymbol{\beta};\boldsymbol{\lambda})$ majorizes $\xi(\boldsymbol{\beta})$ and the MM algorithm derived from $\xi^{[S]}(\boldsymbol{\beta},\boldsymbol{\alpha})$ converges to a stationary point of $\xi(\boldsymbol{\beta})$ if $\xi^{[S]}(\boldsymbol{\beta},\boldsymbol{\alpha})$ is uniquely minimized in $\boldsymbol{\beta}$ for each bounded vector $\boldsymbol{\alpha}$ and at least one of $h(\boldsymbol{\beta},\boldsymbol{\alpha})$ or $q(\boldsymbol{\beta},\boldsymbol{\alpha};\boldsymbol{\lambda}) - p(\boldsymbol{\beta};\boldsymbol{\lambda})$ is strictly positive for each $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$.

For convenience, Table 1 summarizes the various function definitions that will be used throughout the remainder of the paper.

Condition (iii) of Theorem 2.1 establishes convergence under the assumption that $\xi^{[S]}(\boldsymbol{\beta},\boldsymbol{\alpha})$ strictly majorizes $\xi(\boldsymbol{\beta})$ and has a unique minimizer in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha}$. Such a uniqueness condition is shown by Vaida [66] to ensure convergence of the EM and MM algorithms to a stationary point under more restrictive differentiability conditions. Importantly, the assumption of globally strict majorization is only a sufficient condition for convergence; this condition is only important insofar as it guarantees a strict decrease in the objective function at every iteration. It is possible to relax this condition to locally strict majorization, in which $\xi^{[S]}(\boldsymbol{\beta},\boldsymbol{\alpha})$ majorizes $\xi(\boldsymbol{\beta})$, strict majorization being necessary only in an open neighborhood containing $M(\boldsymbol{\alpha})$.

TABLE 1
Function Definitions

Function	Description
$\xi(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + p(\boldsymbol{\beta};\boldsymbol{\lambda}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2$	objective function to be minimized
$g(\boldsymbol{\beta})$	"goodness-of-fit" term
$p(\boldsymbol{\beta};\boldsymbol{\lambda}) = \sum_j \tilde{p}(\lvert\beta_j\rvert;\boldsymbol{\lambda}_j)$	nonsmooth penalty term
$h(\boldsymbol{\beta},\boldsymbol{\alpha})$	function used to majorize $g(\boldsymbol{\beta})$
$q(\boldsymbol{\beta},\boldsymbol{\alpha};\boldsymbol{\lambda}) = \sum_j \tilde{q}(\lvert\beta_j\rvert,\lvert\alpha_j\rvert;\boldsymbol{\lambda}_j)$	function used to majorize $p(\boldsymbol{\beta};\boldsymbol{\lambda})$
$\xi^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \xi(\boldsymbol{\beta}) + h(\boldsymbol{\beta},\boldsymbol{\alpha}) + q(\boldsymbol{\beta},\boldsymbol{\alpha};\boldsymbol{\lambda}) - p(\boldsymbol{\beta};\boldsymbol{\lambda})$	surrogate objective function (majorizer)
$m(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2 + h(\boldsymbol{\beta},\boldsymbol{\alpha})$	smooth portion of $\xi^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\alpha})$
$M(\boldsymbol{\alpha}) = \arg\ \min_{\boldsymbol{\beta}\in\mathbb{R}^p} \xi^{[\mathrm{S}]}(\boldsymbol{\beta},\boldsymbol{\alpha})$	minimization map for MM algorithm

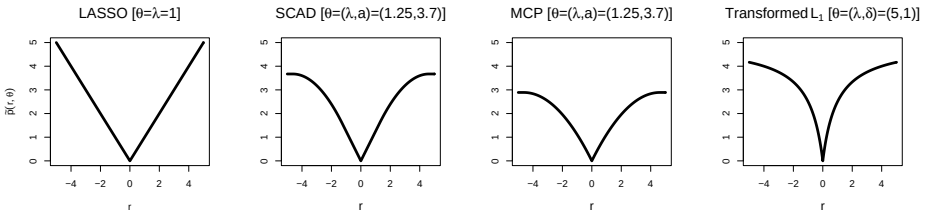

FIG 1. *Four examples of penalties satisfying (P1).*

The use of the MM algorithm and selection of (4) are motivated by the results Zou and Li [77]; we refer the reader to Remark 3.1 below for further comments in this direction. The assumptions on $g(\boldsymbol{\beta})$ clearly cover the case of the linear and canonically parametrized generalized linear models upon setting $g(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$, where $\ell(\boldsymbol{\beta})$ denotes the corresponding loglikelihood function. Other prominent examples include estimation under the semiparametric Cox regression model [14] and accelerated failure time models are also covered upon setting $g(\boldsymbol{\beta})$ to be either the negative logarithm of the partial likelihood function [e.g., 2, Theorem VII.2.1] or the Gehan objective function [e.g., 25, 35].

The assumption (P1) on the penalty function covers a wide variety of popular and interesting examples; see Figure 1 for illustration. For example, the lasso [LAS; e.g., 62], adaptive lasso [ALAS; e.g., 74], elastic net [EN; e.g., 75], and adaptive elastic net [AEN; e.g., 78] penalties are all recovered as special cases upon considering the combination of (3) and the ridge-type penalty $\lambda\varepsilon\|\boldsymbol{\beta}\|^2$. Specifically, with $\boldsymbol{\lambda}_j = (\lambda,\omega_j)^T$ for $\omega_j \geq 0$, taking $\tilde{p}(r;\boldsymbol{\lambda}_j) = \lambda\omega_j r$ in (3) gives LAS ($\omega_j = 1, \varepsilon = 0$), ALAS ($\omega_j > 0, \varepsilon = 0$), EN ($\omega_j = 1, \varepsilon > 0$) and the AEN ($\omega_j > 0, \varepsilon > 0$) penalties. It is easy to see that selecting $\tilde{p}(r;\boldsymbol{\lambda}_j) = \lambda\omega_j r$ also implies the equality of (3) and (4), a result relevant in both (ii) and (iii) of Theorem 2.1 above.

The proposed penalty specification also covers the smoothly clipped absolute deviation [SCAD; e.g., 21] penalty upon setting $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_S(r;\lambda,a)$ for each

$j \geq 1$, where $\tilde{p}_S(r;\lambda,a)$ is defined as the definite integral of

$$\tilde{p}'_S(u;\lambda,a) = \lambda[I(u \leq \lambda) + \frac{(a\lambda - u)_+}{(a-1)\lambda} I(u > \lambda)] \tag{5}$$

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

$$\tilde{p}_S(r;\lambda,a) \leq \tilde{p}_S(s;\lambda,a) + \tilde{p}'_S(s;\lambda,a)(r-s), \tag{6}$$

the boundary cases for $r = 0$ and/or $s = 0$ following from Hiriart-Urruty and Lemaréchal [31, Remark 4.1.2, p. 21]. In other words, $\tilde{p}_S(r;\lambda,a)$ can be majorized by a linear function of r.

The lasso penalty, its variants, and SCAD have received the greatest attention in the literature. More recently, Zhang [71, 72] introduced the minimax concave penalty (MCP), which similarly to SCAD may be defined in terms of its derivative. Specifically, one takes $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_M(r;\lambda,a)$ for each $j \geq 1$ in (3), where $\tilde{p}_M(r;\lambda,a)$ is defined for some fixed $a > 1$ as the definite integral of

$$\tilde{p}'_M(u;\lambda,a) = \left(\lambda - \frac{u}{a}\right)_+ \tag{7}$$

on the interval $0 \leq u \leq r$. Further examples of differentiable concave penalties include the transformed L_1 penalty $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_T(r;\lambda,\delta)$ for

$$\tilde{p}_T(r;\lambda,\delta) = \lambda \frac{\delta r}{1+\delta r}, \ \delta > 0 \tag{8}$$

[e.g., 26, 49]; and $\tilde{p}(r;\boldsymbol{\lambda}_j) = \tilde{p}_Y(r;\lambda,\delta)$ for

$$\tilde{p}_Y(r;\lambda,\delta) = \lambda \log(\delta r + 1), \ \delta > 0; \tag{9}$$

[e.g., 5]. These penalties represent just a small sample of the set of possible penalties satisfying (P1) that one might reasonably consider.

Remark 2.2. *The SCAD and MCP penalties are not strictly concave and lead to surrogate majorizers that fail to satisfy the globally strict majorization condition in (iii) of Theorem 2.1 unless $h(\boldsymbol{\beta},\boldsymbol{\alpha})$ is strictly positive whenever $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$; see Remark 3.1 for further discussion and also Theorem 3.4 below.*

3. Minimization by Iterative Soft Thresholding

3.1. A simple algorithm for convex objective functions

The class of MM algorithms suggested by Theorem 2.1 provides a very general and useful framework for proposing new algorithms in penalized estimation

problems, the key to which is a methodology for solving the minimization problem (1), a step repeated with each iteration of the MM algorithm. Successful application requires the construction of a suitable majorizing function that can be more easily minimized than the desired objective function. In this regard, it is helpful to note that the problem of minimizing $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ for a given $\boldsymbol{\alpha}$ is equivalent to minimizing

$$g(\boldsymbol{\beta}) + \lambda\varepsilon\|\boldsymbol{\beta}\|_2^2 + h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j| \tag{10}$$

in $\boldsymbol{\beta}$. In particular, if $m(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2 + h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is strictly convex for each $\boldsymbol{\alpha}$, which clearly occurs if both $g(\boldsymbol{\beta})$ and $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are convex in $\boldsymbol{\beta}$ and at least one is strictly convex, then (10) is also strictly convex and the corresponding minimization problem has a unique solution.

Remark 3.1. *For $\varepsilon = h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = 0$ and $g(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$ for $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta})$ with $\ell_i(\boldsymbol{\beta})$ a twice continuously differentiable loglikelihood function, the majorizer used by the MM algorithm induced by the surrogate function (10) corresponds (up to sign) to the minorizer employed in the LLA algorithm of Zou and Li [77], an improvement on the so-called LQA algorithm proposed in Hunter and Li [33]. Zou and Li [77, Proposition 1] assert convergence of their LLA algorithm under imprecisely stated assumptions and are additionally unclear as to the nature of the convergence results actually established. For example, while Zou and Li [77, Theorem 1] demonstrate that the LLA algorithm does indeed have an ascent property, their results do not establish the convergence of the LLA solution sequence.*

In contrast, Theorem 2.1 shows that strict majorization, under a few precisely stated conditions, is sufficient to ensure local convergence of the resulting MM algorithm to a stationary point of (2). In Section 3.2, it is further demonstrated how a particular choice of $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ yields a strict majorizer that permits both closed form minimization and componentwise updating at each step of the MM algorithm, even in the case of penalties that fail to be strictly concave.

Numerous methods exist for minimizing a differentiable convex objective function [e.g., 10]. However, because (10) is not differentiable, such methods do not apply in the current setting. Specialized methods exist for nonsmooth problems of the form (10) in settings where $g(\boldsymbol{\beta})$ has a special structure; a well-known example here is LARS [19], which can be used to efficiently solve lasso-type problems in the case where $g(\boldsymbol{\beta})$ is replaced by a least squares objective function. Recently, Combettes and Wajs [13, Proposition 3.1; Theorem 3.4] proposed a general class of fixed point algorithms for minimizing $f_1(h) + f_2(h)$, where $f_i(\cdot), i = 1, 2$ are each convex and h takes values in some real Hilbert space $\mathcal{H}$. Hale, Yin and Zhang [29, Theorem 4.5] specialize the results of Combettes and Wajs [13] to the case where $\mathcal{H}$ is some subset of $\mathbb{R}^p$ and $f_2(h) = \sum_{j=1}^{p} |h_i|$. The collective application of these results to the problem of minimizing (10) generates an iterated soft-thresholding procedure with an appealingly simple structure. Theorem 3.2, given below, states the algorithm along with conditions

under which the algorithm is guaranteed to converge; a proof is provided in Appendix A.3. The resulting class of procedures, that is, MM algorithms in which the minimization of (10) is carried out via iterated soft-thresholding, is hereafter referred to as MIST, an acronym for (M)inimization by (I)terated (S)oft (T)hresholding. Two important and useful features of MIST include the absence of high-dimensional matrix inversion and the ability to update each individual parameter separately.

Theorem 3.2. *Let $p(\cdot;\boldsymbol{\lambda})$ satisfy the assumptions of Section 2.2. Suppose $m(\boldsymbol{\beta})$ in Table 1 is strictly convex with a Lipschitz continuous derivative of order $L^{-1} > 0$ for each bounded vector $\boldsymbol{\alpha}$. Then, for any such $\boldsymbol{\alpha}$ and a constant $\varpi \in (0, 2L)$, the unique minimizer of (10) can be obtained in a finite number of iterations using iterated soft-thresholding:*

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

$$S(\mathbf{u};\mathbf{v}) = \sum_{j=1}^{p} s(u_j, v_j)\,\mathbf{e}_j, \tag{11}$$

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

$$s(u_j, v_j) = sign(u_j)(|u_j| - v_j)_+, \tag{12}$$

is the univariate soft-thresholding operator, and $\boldsymbol{\tau} = (\tilde{p}'(|\alpha_1|;\boldsymbol{\lambda}_1), \ldots, \tilde{p}'(|\alpha_p|;\boldsymbol{\lambda}_p))^T$.
3. *Stop if converged; else, set $n = n + 1$ and return to Step 2.*

Theorem 3.4 of Combettes and Wajs [13] shows that the update in Step 2 can be generalized to

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

where $\varpi_n \in (0, 2L)$ and $\delta_n \in (0, 1]$ is a suitable sequence of relaxation constants. Judicious selection of these constants, possibly updated at each step, can improve the convergence rate. In principle, the minimization algorithm of Theorem 3.2 can also be replaced with any other suitable minimization algorithm. For example, since the penalty term appearing in (10) is in fact a separable convex function of the parameters, one could instead employ the coordinate gradient descent method recently proposed in Tseng and Yun [65]. An advantage of the proposed approach is its computational simplicity; moreover, as will be seen in Section 3.2, the proposed soft-thresholding update arises naturally in a wide class of minimization problems of interest to statisticians.

Theorem 3.2 imposes the condition that the gradient of $m(\boldsymbol{\beta})$ is globally L^{-1}-Lipschitz continuous. The role of this condition, also imposed in Combettes and Wajs [13, Proposition 3.1; Theorem 3.4], is to ensure that the update at each step of the proposed algorithm is a contraction, thereby guaranteeing its convergence

to a fixed point; see Schifano [57] for a proof of this result. However, in view of the generality of the Contraction Mapping Theorem [e.g., 42, Theorem 10.2.1], it is possible to relax the requirements on $\nabla m(\boldsymbol{\beta})$ provided that one selects a suitable starting point. A useful extension is summarized in the corollary below; one may prove this result in a manner similar to Theorem 4.5 of Hale, Yin and Zhang [29]. As a reminder to the reader, the relevant optimization problem at this stage involves a specified vector $\boldsymbol{\alpha}$ having bounded elements.

Corollary 3.3. *Let $p(\cdot;\boldsymbol{\lambda})$ satisfy the assumptions of Section 2.2 and let $\boldsymbol{\alpha}$ be given. Suppose $m(\boldsymbol{\beta})$ is strictly convex and twice continuously differentiable function of $\boldsymbol{\beta} \in \Omega$, where $\Omega \subset \mathbb{R}^p$ is a convex, compact set. Then, there exists a unique minimizer $\boldsymbol{\beta}^*$ of (10) on Ω and the algorithm of Theorem 3.2 converges to $\boldsymbol{\beta}^*$ in a finite number of iterations provided that $\mathbf{b}^{(0)} \in \Omega$, $\lambda^* = \max_{\boldsymbol{\beta}\in\Omega} \lambda_{max}(\boldsymbol{\beta}) < \infty$ and $\varpi \in (0, 2/\lambda^*)$, where $\lambda_{max}(\boldsymbol{\beta})$ denotes the maximum eigenvalue of $\nabla^2 m(\boldsymbol{\beta})$.*

Some useful insight into the form of the proposed thresholding algorithm can be gained by considering the behavior of the penalty derivative term $\tilde{p}'(r;\boldsymbol{\theta})$. As suggested earlier, (P1) implies that $\tilde{p}'(r;\boldsymbol{\theta})$ decreases from its maximum value towards zero as r moves away from the origin. For some penalties (e.g., SCAD, MCP), this derivative actually becomes zero at some finite value of $r > 0$, resulting in situations for which $\tau_j = \tilde{p}'(|\alpha_j|;\boldsymbol{\lambda}_j) = 0$ for at least one j. If this occurs for component j, then j^{th} component of the vector $S\big(\mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)}); \varpi\boldsymbol{\tau}\big)$ simply reduces to the j^{th} component of the argument vector $\mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)})$. In the extreme case where $\boldsymbol{\tau} = \mathbf{0}$, the proposed update reduces to $\mathbf{b}^{(n+1)} = \mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)})$, a steepest descent step; equivalently, the algorithm takes an inexact Newton step in which the inverse hessian matrix is replaced by $\varpi\mathbf{I}_p$, $\mathbf{I}_p$ denoting the $p \times p$ identity matrix, and with step-size chosen to ensure that this update yields a contraction. Hence, if each of the components of $\mathbf{b}^{(n)} - \varpi\nabla m(\mathbf{b}^{(n)})$ are sufficiently large in magnitude, the proposed algorithm simply takes an inexact Newton step towards the solution; otherwise, one or more components of this vector may be thresholded. Notably, replacing $\varpi\mathbf{I}_p$ with any diagonal matrix having bounded entries preserves the componentwise nature of the proposed algorithm; alternative strategies that both adapt the step size to each component and maintain the indicated convergence properties are worthy of further investigation.

3.2. Penalized estimation for generalized linear models

The combination of Theorems 2.1, 3.2 and Corollary 3.3 lead to a useful and stable class of algorithms with the ability to deal with a wide range of penalized regression problems. In settings where $g(\boldsymbol{\beta})$ is strictly convex and twice continuously differentiable, one can safely assume that $h(\boldsymbol{\beta},\boldsymbol{\alpha}) = 0$ for all choices of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ provided that $\tilde{p}'(r;\boldsymbol{\theta})$ in (P1) is strictly positive for $r > 0$; important examples of statistical estimation problems here include many commonly used linear and generalized linear regression models, semiparametric Cox regression

[14], and smoothed versions of the accelerated failure time regression model [cf. 35]. The SCAD and MCP penalizations, as well as other penalties having $\tilde{p}'(r;\boldsymbol{\theta}) \geq 0$ for $r > 0$, can also be used; however, additional care is required. In particular, and as pointed out in an earlier remark, if one sets $h(\boldsymbol{\beta},\boldsymbol{\alpha}) = 0$ for all $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ then convergence of the resulting algorithm to a stationary point is no longer guaranteed by the above results due to the resulting failure of these penalties to induce strict majorization.

The need to use an iterative algorithm for repeatedly minimizing (10) is not unusual for the class of MM algorithms. However, it turns out that for certain choices of $g(\boldsymbol{\beta})$, a suitable choice of $h(\boldsymbol{\beta},\boldsymbol{\alpha})$ in Theorem 3.2 guarantees both strict majorization and permits one to compute the minimizer of (10) in closed form (i.e., in one step), resulting in a single soft-thresholding update at each iteration. Below, we demonstrate how the MIST algorithm simplifies in settings where $g(\boldsymbol{\beta})$ corresponds to the negative loglikelihood function of a canonically parametrized generalized linear regression model with a uniformly bounded hessian matrix. The result applies to all penalties satisfying condition (P1), including SCAD and MCP. A proof is provided in Appendix A.4.

Theorem 3.4. *Let $\boldsymbol{y}$ be $N \times 1$ and suppose the probability distribution of $\boldsymbol{y}$ follows a generalized linear model with a canonical link and linear predictor $\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}$, where $\widetilde{\boldsymbol{X}} = [\boldsymbol{1}_N, \boldsymbol{X}]$ is $N \times (p+1)$ and $\widetilde{\boldsymbol{\beta}} = [\beta_0, \boldsymbol{\beta}^T]^T$ is $(p+1) \times 1$ with β_0 denoting an intercept. Define $g(\widetilde{\boldsymbol{\beta}}) = -\ell(\widetilde{\boldsymbol{\beta}})$, where*

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

is the corresponding loglikelihood, $\widetilde{\boldsymbol{\eta}} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}$ with elements $\tilde{\eta}_i$ and $E[y_i] = d'(\tilde{\eta}_i)$, $i = 1, \ldots, N$, for $d(\cdot)$ strictly convex and twice continuously differentiable. Let $\lambda_{max}(\widetilde{\boldsymbol{\beta}})$ denote the largest eigenvalue of $-\nabla^2\ell(\widetilde{\boldsymbol{\beta}})$ and assume that $\lambda^* = \max_{\widetilde{\boldsymbol{\beta}}} \lambda_{max}(\widetilde{\boldsymbol{\beta}}) < \infty$.

Let the penalty function be defined as in (3) and satisfy (P1); note that β_0 remains unpenalized. Define

$$h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = \ell(\widetilde{\boldsymbol{\beta}}) - \ell(\widetilde{\boldsymbol{\alpha}}) - \nabla\ell(\widetilde{\boldsymbol{\alpha}})^T(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\alpha}}) + \varpi^{-1}(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\alpha}})^T(\widetilde{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\alpha}}); \tag{13}$$

where $\tilde{\boldsymbol{\alpha}} \equiv [\alpha_0, \boldsymbol{\alpha}^T]^T$ is $(p+1) \times 1$, and $\varpi \in (0, 2/\lambda^*)$. Finally, suppose $\xi(\widetilde{\boldsymbol{\beta}})$, defined in (2), satisfies $\xi(\widetilde{\boldsymbol{\beta}}) > -\infty$ for $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$ and that its corresponding set of stationary points, defined in the sense of Clarke [12], is non-empty, finite and consists only of bounded local or global minima. Then:

1. *Up to a constant independent of $\widetilde{\boldsymbol{\beta}}$, objective function $\xi(\tilde{\boldsymbol{\beta}})$ in (2) is majorized by*

$$\xi^{[S]}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}) = -\ell(\tilde{\boldsymbol{\alpha}}) - \nabla\ell(\tilde{\boldsymbol{\alpha}})^T(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\alpha}}) + \varpi^{-1}(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\alpha}})^T(\tilde{\boldsymbol{\beta}} - \tilde{\boldsymbol{\alpha}}) + \sum_{j=1}^{p} (\tau_j|\beta_j| + \lambda\varepsilon\beta_j^2) \tag{14}$$

   where $\tau_j = \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)$.

2. *The functions* $g(\widetilde{\boldsymbol{\beta}}) = -\ell(\tilde{\boldsymbol{\beta}})$, $p(\boldsymbol{\beta};\boldsymbol{\lambda})$ *and* $h(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})$ *satisfy the regularity conditions of Theorem 2.1; hence, an MM algorithm that uses* (14) *converges to a minimizer of* (2).
3. *For each bounded vector* $\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^{p+1}$,
   (a) *the minimizer* $\tilde{\boldsymbol{\beta}}^*$ *of* $\xi^{[S]}(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}})$ *is unique and satisfies*

$$\begin{aligned} \boldsymbol{\beta}^* &= \frac{1}{1+\varpi\lambda\varepsilon} S\left(\boldsymbol{\alpha} + \frac{\varpi}{2}[\nabla\ell(\tilde{\boldsymbol{\alpha}})]_{\mathcal{A}}, \frac{\varpi}{2}\boldsymbol{\tau}\right), \\ \beta_0^* &= \alpha_0 + \frac{\varpi}{2}[\nabla\ell(\tilde{\boldsymbol{\alpha}})]_0 \end{aligned} \tag{15}$$

   where $S(\cdot;\cdot)$ *is the soft-thresholding operator defined in* (11) *and* $\mathcal{A} = \{1,\ldots,p\}$.
   (b) *for each bounded vector* $\tilde{\boldsymbol{\kappa}} \equiv [\kappa_0, \boldsymbol{\kappa}^T]^T \in \mathbb{R}^{(p+1)}$,

$$\xi^{[S]}(\tilde{\boldsymbol{\beta}}^* + \tilde{\boldsymbol{\kappa}}, \tilde{\boldsymbol{\alpha}}) \geq \xi^{[S]}(\tilde{\boldsymbol{\beta}}^*, \tilde{\boldsymbol{\alpha}}) + \varpi^{-1}\left\|\tilde{\boldsymbol{\kappa}}\right\|^2. \tag{16}$$

In view of previous results, the result in # 3 of Theorem 3.4 shows that the resulting MM algorithm takes a very simple form: given the current iterate $\widetilde{\boldsymbol{\beta}}^{(n)}$,

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

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

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

$$\boldsymbol{\beta}^{(n+1)} = \frac{1}{1+\varpi\lambda\varepsilon} S\left(\boldsymbol{\beta}^{(n)} + \frac{\varpi}{2}[\nabla\ell(\tilde{\boldsymbol{\beta}}^{(n)})]_{\mathcal{A}}; \frac{\varpi}{2}\boldsymbol{\tau}^{(n)}\right), \tag{17}$$

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

The proposed algorithm can be easily generalized to accommodate additional regression variables (besides the intercept) not subject to penalization. The specific choice of function $h(\widetilde{\boldsymbol{\beta}},\widetilde{\boldsymbol{\alpha}})$ clearly serves two useful purposes: (i) it leads to componentwise-soft thresholding; and, (ii) it leads to strict majorization, as is required in condition (iii) of Theorem 2.1, allowing one to establish the convergence of MIST for SCAD and other penalties having $\tilde{p}'(r,\boldsymbol{\theta}) = 0$ at some finite $r > 0$.

An important class of problems to which these results apply is the setting of penalized linear regression. Suppose that $\mathbf{y}$ has been centered to remove β_0 from consideration and that the problem has also been rescaled so that $\mathbf{X}$, which is now $N \times p$, satisfies the indicated conditions. Then, Theorem 3.4 applies directly with

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

where ϖ is defined as in Theorem 3.4 with λ^* set equal to the largest eigenvalue of $\mathbf{X}'\mathbf{X}$. For the class of adaptive elastic net penalties (i.e., $\tilde{p}(r;\boldsymbol{\lambda}_j) = \lambda\omega_j r$ in (3)), the resulting iterative scheme is exactly that proposed in De Mol, De Vito and Rosasco [17, pg. 17], specialized to the setting of a Euclidean parameter. In particular, we have $\tau_j = \lambda\omega_j$ and $\gamma_j = 0$ in Theorem 3.4, and the proposed update reduces to

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

where $\nu = 2\varpi^{-1}$. Setting $\nu = 1$ and $\varepsilon = 0$ yields the iterative procedure proposed in Daubechies, Defreise and De Mol [16] provided $\mathbf{X}'\mathbf{X}$ is scaled such that $\mathbf{I} - \mathbf{X}'\mathbf{X}$ is positive definite. The MIST algorithm extends these iterative soft-thresholding procedures to a much wider class of penalized estimation problems.

In an interesting unpublished paper, Mazumder, Friedman and Hastie [45] propose the SparseNet algorithm, a coordinatewise descent algorithm for minimizing objective functions of the form (2) with $g(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, $\varepsilon = 0$ and $p(\boldsymbol{\beta};\boldsymbol{\lambda})$ a family of penalty functions satisfying (3) and several additional regularity conditions. Their specification includes the lasso, SCAD and MCP penalties, as well as several other examples of nonconvex penalties. The full SparseNet algorithm intends to generate the solution *surface* as a function of the penalty parameter λ and a parameter γ indexing the penalty family (i.e., restricted to a two dimensional grid). While the algorithm incorporates a number of useful features, solutions are found for each (λ, γ) pair using a simple coordinate descent algorithm. In the case of the lasso penalty ($\gamma = \infty$) and provided $\mathbf{X}$ is column-standardized, this coordinate descent algorithm is almost identical to the componentwise soft-thresholding algorithm proposed in Daubechies, Defreise and De Mol [16] (hence MIST), the primary differences stemming from the form of the iterative update (i.e., the use of a simultaneous update implemented via componentwise soft-thresholding versus cyclical application of the soft-thresholding operator). For other penalties, such as SCAD and MCP, the coordinatewise updates utilized by SparseNet rely on so-called generalized thresholding operators [cf. 59], departing more substantially from the iterated soft-thresholding procedure used in the MIST algorithm. Mazumder, Friedman and Hastie [45] provide an explicit proof of the convergence of the solution sequence obtained for a given (λ, γ) pair. The regularity conditions under which these results are obtained appear to be similarly weak to those required by Theorem 3.4 (i.e., applied to the penalized least squares problem). However, unlike MIST, it not obvious how to extend the SparseNet algorithm to more general choices of $g(\boldsymbol{\beta})$ in the absence of reparameterizations that permit componentwise separation of parameters.

The restriction to canonical generalized linear models in Theorem 3.4 is imposed to ensure strict convexity of the negative loglikelihood. Our results are easily modified to handle non-canonical generalized linear models, provided the negative loglikelihood remains strictly convex in $\widetilde{\boldsymbol{\beta}}$ and the hessian can be appropriately bounded. Interestingly, not all canonically parametrized generalized linear models satisfy the regularity conditions of Theorem 3.4. For

example, in the classical setting of N independent Poisson observations with $E[Y_i|\tilde{X}_i] = \mathcal{O}_i \exp\{\tilde{x}_i^T \widetilde{\boldsymbol{\beta}}\}$ for a known set of scalar (log) offsets $\mathcal{O}_1, \ldots, \mathcal{O}_N$, we have (i.e., up to irrelevant constants) $\ell(\widetilde{\boldsymbol{\beta}}) = -\sum_{i=1}^N f_i(\tilde{x}_i^T \widetilde{\boldsymbol{\beta}})$, where

$$f_i(u) = \mathcal{O}_i e^u - y_i u.$$

It is easy to see that $\nabla \ell(\widetilde{\boldsymbol{\beta}})$, hence $\nabla m(\widetilde{\boldsymbol{\beta}})$, is locally but not globally Lipschitz continuous; hence, it is not possible to choose a matrix $\mathbf{C} = \varpi^{-1}\mathbf{I}$ such that (14) everywhere majorizes $\xi(\widetilde{\boldsymbol{\beta}})$. Nevertheless, progress remains possible. For example, Corollary 3.3 implies that one can still use a single update of the form (17) provided that a suitable Ω, hence $\mathbf{C}$ and $\widetilde{\boldsymbol{\beta}}^{(0)}$, can be identified. Alternatively, using results summarized in Becker, Yang and Lange [6], one can instead majorize $-\ell(\widetilde{\boldsymbol{\beta}})$ for any bounded $\widetilde{\boldsymbol{\alpha}}$ using

$$k(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = \sum_{j=0}^p k_j(\beta_j; \alpha_j)$$

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

where, for every i, $\theta_{ij} \geq 0$ are any sequence of constants satisfying $\sum_{j=0}^p \theta_{ij} = 1$ and $\theta_{ij} > 0$ if $x_{ij} \neq 0$. Of importance here is the fact $k_j(\beta_j; \alpha_j)$ is a strictly convex function of β_j and does not depend on β_k for $k \neq j$. One may now take $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ in Theorem 2.1 as being equal to $k(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) + \ell(\widetilde{\boldsymbol{\beta}})$, leading to the minimization of

$$\xi^{[\mathrm{S}]}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) \propto \sum_{j=1}^p [k_j(\beta_j; \alpha_j) + \lambda\varepsilon\beta_j^2 + \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|] + k_0(\beta_0, \alpha_0). \tag{18}$$

In particular, componentwise soft-thresholding is replaced by componentwise minimization of (18), the latter using any algorithm capable of minimizing a continuous univariate convex function.

Remark 3.5. *The Cox proportional hazards model [14], while not a generalized linear model, shares the essential features of the generalized linear model utilized in Theorem 3.4. In particular, the negative log partial likelihood, say* $g(\boldsymbol{\beta}) = -\ell_p(\boldsymbol{\beta})$, *is (under mild regularity conditions) strictly convex, twice continuously differentiable, and has a bounded hessian [e.g., 2, 9]. Consequently, Theorem 3.4 and its proof are easily modified for this setting upon taking* $g(\boldsymbol{\beta})$ *as indicated, setting* $h(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \ell_p(\boldsymbol{\beta}) - \ell_p(\boldsymbol{\alpha}) - \nabla\ell_p(\boldsymbol{\alpha})^T(\boldsymbol{\beta} - \boldsymbol{\alpha}) + \varpi^{-1}\|\boldsymbol{\beta} - \boldsymbol{\alpha}\|^2$, *and taking* $\varpi \in (0, 2/\lambda^*)$ *where* $\lambda^* = \max_{\boldsymbol{\beta}} \lambda_{max}(\boldsymbol{\beta}) < \infty$ *where* $\lambda_{max}(\boldsymbol{\beta})$ *is the largest eigenvalue of* $-\nabla^2\ell_p(\boldsymbol{\beta})$.

3.3. Accelerating convergence

Similarly to the EM algorithm, the stability and simplicity of the MM algorithm frequently comes at the price of an increased number of iterations before convergence. Numerous methods of accelerating the EM algorithm have

been proposed in the literature; see McLachlan and Krishnan [46] for a review. Recently, Varadhan and Roland [67] proposed a new method for EM called SQUAREM, obtained by "squaring" an iterative Steffensen-type (STEM) acceleration method. Both STEM and SQUAREM are structured for use with iterative mappings of the form $\theta_{n+1} = M(\theta_n)$, $n = 0, 1, 2, \ldots$, hence applicable to both the EM and MM algorithms. Specifically, the acceleration update for SQUAREM is given by

$$\begin{aligned} \theta_{n+1} &= \theta_n - 2\gamma_n(M(\theta_n) - \theta_n) + \gamma_n^2[M(M(\theta_n)) - 2M(\theta_n) + \theta_n] \\ &= \theta_n - 2\gamma_n r_n + \gamma_n^2 v_n, \end{aligned} \tag{19}$$

where $r_n = M(\theta_n) - \theta_n$ and $v_n = (M(M(\theta_n)) - M(\theta_n)) - r_n$ for an adaptive steplength γ_n. Varadhan and Roland [67] suggest several steplength options, with preference for the choice $\gamma_n = -\|r_n\|/\|v_n\|$. Roland and Varadhan [53] provide a proof of local convergence for SQUAREM under restrictive conditions on the EM mapping $M(\theta)$, while Varadhan and Roland [67] outline a proof for global convergence for versions of SQUAREM that employ a back-tracking strategy. We study the effectiveness of SQUAREM applied to the simplified form of the MIST algorithm, hereafter denoted SQUAREM^2, in Section 4.3.

4. Simulation results

The simulation results summarized below are intended to compare the estimates of $\boldsymbol{\beta}$ obtained from existing methods to those obtained using the simplified MIST algorithm of Theorem 3.4. In particular, we consider the performance of MIST for the class of penalized linear and generalized linear models, demonstrating its capability of recovering the solutions provided by existing algorithms when both algorithms are forced to use the same set of "tuning" parameters (i.e., penalty parameter(s), plus any additional parameters required to define the penalty itself). In cases where multiple local minima can arise, we further show that the MIST algorithm often tends to find solutions with lower objective function evaluations in comparison with existing algorithms, provided these algorithms utilize the same choice of starting value.

4.1. Example 1: Linear model

Let $\mathbf{1}_m$ and $\mathbf{0}_m$ respectively denote m-dimensional vectors of ones and zeros. Then, following Zou and Zhang [78], we generated data from the linear regression model

$$y = \mathbf{x}'\boldsymbol{\beta}^* + \epsilon \tag{20}$$

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

Penalized least squares estimation is considered for five popular choices of penalty functions, all of which are currently implemented in the R software language: LAS, ALAS, EN, AEN, and SCAD. The LAS, ALAS, EN and AEN penalties are all convex and lead to unique solutions under mild conditions; the SCAD penalty is concave and the resulting minimization problem is, depending on the design matrix and choice of a, either convex or nonconvex [cf. 72]. The SCAD examples considered here lead to non-convex objective functions, hence may have multiple solutions. In each case, we used existing software for computing solutions subject to these penalizations and compared those results to the solutions computed using the MIST algorithm. For the MIST algorithm, ϖ was chosen to be $2/(\lambda^* + .001)$ where λ^* is the largest eigenvalue of $\mathbf{X}'\mathbf{X}$ where $\mathbf{X}$ is appropriately scaled to match the scaling of the existing algorithm.

Regarding existing methods, we respectively used the *lars* [30] and *elasticnet* [76] packages for computing solutions in the case of the LAS and EN penalties. For the ALAS and AEN penalties, we used software kindly provided by Zou and Zhang [78] that also makes use of the *elasticnet* package. The weights for the AEN penalty are computed using $\omega_j = |\hat{\beta}_j^{EN}|^{-\gamma}$, $j = 1, \ldots, p$, where $\hat{\boldsymbol{\beta}}^{EN}$ is an EN estimator and γ is a positive constant. Using EN-based weights in the AEN fitting algorithm necessitates tuning parameter specification for both EN and AEN. As in Zou and Zhang [78], the ℓ_1 parameters λ (λ_1 in their notation) are allowed to differ, whereas the ℓ_2 parameters ε (λ_2 in their notation) are forced to be the same. Evidently, setting $\varepsilon = 0$ ($\lambda_2 = 0$) results in the ALAS solution. For the SCAD penalty, we considered the estimator of Kim, Choi and Oh [36] (HD), as well the one-step SCAD (1S) and LLA estimators of Zou and Li [77]. The code for the first two methods was kindly provided by their respective authors; the LLA estimator was computed using the R package *SIS*. The choice $a = 3.7$ was used for all implementations of SCAD.

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

For the strictly convex objective functions associated with the LAS, ALAS, EN, and AEN penalties, we simply used a starting value of $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p$. For SCAD, three different starting values for the MIST, HD, and LLA SCAD algorithms were considered: $\boldsymbol{\beta}^{(0)} = \mathbf{0}_p$, $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}_{ml}$ (i.e., the unpenalized least squares estimate), and $\boldsymbol{\beta}^{(0)} = \widehat{\boldsymbol{\beta}}_{1S,\lambda}$ (i.e., the one-step estimate computed using the penalty parameter λ). As in Zou and Li [77], the one-step estimator is computed using $\widehat{\boldsymbol{\beta}}_{ml}$, an appropriate choice when $N > p$.

The convergence criteria used by the existing software packages were used without alteration in this simulation study. The convergence criteria used for

TABLE 2
Maximum average normed differences ($\times 10^5$) over $B = 100$ simulations for Examples 1 (Linear Model; LM) and 2 (Generalized Linear Model; GLM)

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

LAS = Lasso; ALAS = Adaptive Lasso; EN = Elastic Net; AEN = Adaptive Elastic Net

MIST were as follows: the algorithm stopped if either (i) the normed difference of successive iterates was less than 10^{-6} (convergence of coefficients); or, (ii) the difference of the objective function evaluated at successive iterates was less than 10^{-6} and the number of iterations exceeded 10^6 (convergence of optimization). Due to the large number of comparisons and highly intensive nature of the computations, we ran $B = 100$ simulations for each choice of ρ, σ, and p. We report the results for the convex penalties in Table 2 and those for the SCAD penalty in Tables 3 and 4.

In Table 2, we summarize the average normed difference between the solution obtained using existing software and that obtained using MIST, $\|\hat{\boldsymbol{\beta}}_{exist} - \hat{\boldsymbol{\beta}}_{mist}\|$, over the $B = 100$ simulations; in particular, we report in the two leftmost panels the maximum value of this difference, computed across all combinations of tuning parameters. These maximum differences (all of which are multiplied by 10^5) are remarkably small for all (A)LAS and (A)EN penalties, indicating that MIST recovers the same (unique) solutions as the existing algorithms. Interestingly, the values for LAS are slightly larger than the rest, where the maximum differences all resulted from the smallest value of λ considered ($\lambda = 0.1$). In these cases, the algorithm tended to stop using the objective function criteria rather than the (stricter) coefficient criteria, resulting in slightly larger differences on average.

The results for SCAD are reported in Tables 3 ($p = 35$) and 4 ($p = 81$) and display (i) the average normed differences, multiplied by 10^3, for each combination of λ, ρ, σ, p and starting value; and, (ii) the proportion of simulated datasets in which the MIST solution yields a lower or equivalent evaluation of the objective function in comparison with the solution obtained using another method for the indicated choice of starting value. We remark here that SCAD penalties used in the existing implementations are multiplied by a factor of N, i.e., $p(\boldsymbol{\beta}; \boldsymbol{\lambda}) = \sum_{j=1}^{p} N \tilde{p}_S(|\beta_j|; \lambda, a)$, so the MIST implementation incorporates this factor of N as well. The results for $\lambda = 100$ are not shown, as the solution was $\mathbf{0}_p$ in all cases. In comparison with the convex penalties, larger normed

TABLE 3

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

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

HD=High Dimensional SCAD [36]; 1S=one-step & LLA=local linear approximation [77]

differences are observed, even when controlling for the use of the same starting value. Such differences are a result of two important features of the SCAD optimization problem: (i) the possible existence of several local minima; and, (ii) the fact that the MIST, HD, and LLA algorithms each take a different path from a given starting value towards one of these solutions. For example, while each of the LLA, MIST, and HD algorithms involve majorization of the objective function using a lasso-type surrogate objective function, both the majorization

TABLE 4

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

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

HD=High Dimensional SCAD [36]; 1S=one-step & LLA=local linear approximation [77]

and minimization of the resulting surrogate function are carried out differently in each case. In particular, the LLA algorithm, as implemented in *SIS*, majorizes only the penalty term and adapts the lasso code in *glmpath* in order to minimize the corresponding surrogate objective function at each step. The HD algorithm is similar in spirit, but instead decomposes the penalty term into a sum of a concave and convex function and utilizes the the algorithm of Rosset and Zhu [55] to minimize the corresponding surrogate objective function. The MIST al-

gorithm instead uses the same penalty majorization as the LLA algorithm, but additionally majorizes the negative loglikelihood term in a way that permits minimization of the surrogate function in a single soft-thresholding step. Each procedure therefore takes a different path towards a solution, even when given the same starting value.

We remark here that differences must also expected between any of LLA, HD, MIST and 1S. From an optimization perspective, the one-step estimator is the result of running just one iteration of the LLA algorithm, starting from the unpenalized least squares estimator $\widehat{\boldsymbol{\beta}}_{ml}$ [77]; hence, 1S only provides an approximate solution to the desired minimization problem. All other methods (LLA, MIST, HD) iterate until some local minimizer (or stationary point) is reached. For example, when using either $\widehat{\boldsymbol{\beta}}_{ml}$ or $\widehat{\boldsymbol{\beta}}_{1S,\lambda}$ as the starting value, MIST always found a solution that produced a lower evaluation of the objective function in comparison to $\widehat{\boldsymbol{\beta}}_{1S,\lambda}$. However, when using the null starting value of $\mathbf{0}_p$, the one-step estimator did occasionally result in a lower objective function evaluation in cases involving smaller values of λ. This behavior is not terribly surprising; with small λ, the one-step solution should generally be close to the unpenalized least squares solution, as the objective function itself is likely to be dominated by the least squares term.

Of all the SCAD algorithms considered here, MIST and LLA tended to find the most similar solutions (i.e., have the smallest normed differences). For the cases in which the LLA solution had lower objective function evaluations, all of the MIST solutions were also LLA solutions; i.e., when starting the LLA algorithm with the MIST solution, the algorithm terminated at the starting value (i.e., the LLA solution coincides with the MIST solution). With the exception of three of these cases, starting the MIST algorithm with the LLA solution also resulted in the same behavior. The HD and MIST algorithms also generally gave similar results, with one source of difference being the respective stopping criteria used. The stopping criteria for HD, assessed in order, are as follows: (1) 'convergence of optimization': stop if the absolute value of the difference of the objective evaluated at successive iterates is less than 10^{-6}; (2) 'convergence of penalty gradient': stop if the sum of the absolute value of the differences of the derivative of the centered penalty evaluated at successive iterates is less than 10^{-6}; (3) 'convergence of coefficients:' stop if the sum of the absolute value of the differences of successive iterates is less than 10^{-6}; and, (4) 'jump-over' criteria: stop if the objective at the previous iterate plus 10^{-6} was less than the objective at the current iterate. After careful analysis of the results, we assert the following:

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

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

4.2. Example 2: Binary logistic regression

As in Example 1, we considered the LAS, ALAS, EN, AEN, and SCAD penalties. There are at least two R packages that allow penalization using the LAS and EN penalties: *glmpath* [51], which handles binomial and poisson regression using a "predictor-corrector" method, and *glmnet* [24], which handles binomial and multinomial regression using cyclical coordinate descent. Both methods can be tuned to find the same solutions, so for ease of presentation we only consider the results of *glmnet* for comparison in the tables and analysis below. The *SIS* package [22] permits computations with the ALAS, AEN, and SCAD penalties using modifications of the Park and Hastie [51] code. For SCAD, we compared the results of MIST to the results from the one-step (1S) algorithm [GLM version, 77] using the code provided from the authors and the LLA algorithm as implemented in Fan *et al.* [22].

As before, we only considered comparing solutions that use the same combination of tuning parameters; for the present example, the set considered here is $\Lambda = \{0.001, 0.01, 0.05, 0.1, 0.2, 1.00\}$, reflecting a need to accommodate the different scaling of the problem. The data generation scheme for this example was loosely based on the simulation study found in Friedman, Hastie and Tibshirani [24]. Binary response data were generated according to a logistic (rather than linear) regression model using $p_i = [1+\exp(-\mathbf{x}_i'\boldsymbol{\beta}^*)]^{-1}$, $i = 1, \ldots, N = 1000$, where $\boldsymbol{\beta}^*$ is a $p-$vector with elements $\beta_j = 3 \times (-1)^j \exp(-2(j-1)/200)$, $j = 1, \ldots, q$, $q \in \{25, 75\}$, and the remaining $100 - q$ components set to zero. Here, $\mathbf{x}_i$ follows a p-dimensional multivariate normal distribution with zero mean and covariance $\boldsymbol{\Sigma} = 3^{-2}\mathbf{P}$ where the correlation matrix $\mathbf{P}$ is such that each pair of predictors has the same population correlation ρ. We considered three such correlations, $\rho \in \{0.0, 0.5, 0.75\}$. For the MIST algorithm, ϖ was selected to be $2/(\lambda^* + 0.001)$ where λ^* is the largest eigenvalue of $\mathbf{X}'\mathbf{W}\mathbf{X}$ where $\mathbf{W} = .25\mathbf{I}_N$ where the design matrix $\mathbf{X}$ is appropriately scaled to match the scaling of the existing algorithm.

TABLE 5

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

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

1S=one-step & LLA=local linear approximation [77]

For the $B = 100$ simulations, the maximum (across different tuning parameters) average normed difference of solutions between the existing and proposed methods, multiplied by 10^5, are reported for each of the strictly convex cases in the right-most panel of Table 2. As before, these maximums are generally remarkably small, indicating that MIST recovers the same (unique) solutions as the existing algorithms. The results for SCAD are reported in Table 5, which displays the same information as in the corresponding tables from Example 1; the HD comparisons are omitted here as the methodology and code were only developed for the case of penalized least-squares. In the GLM setting, the 1S estimator is computed by applying the LARS [19] algorithm to a quadratic approximation of the negative loglikelihood function evaluated at the MLE. In contrast, both MIST and LLA utilize the exact objective function and iterate

until a stationary point, usually a local minimizer, is found. As in the linear model case, LLA uses *glmpath* to minimize the surrogate at each step, whereas the MIST algorithm uses a single application of the soft thresholding operator to minimize the surrogate function at each step.

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

There were a few settings in which the 1S estimator resulted in a lower objective function evaluation in comparison with applying MIST started at $\widehat{\boldsymbol{\beta}}_{ml}$. This reflects the fact that several local minima can exist for non-convex penalties like SCAD. In addition, and as was observed before, using the 1S solution as a starting value always led to MIST finding a solution with a lower evaluation of the objective function in comparison with that provided by the 1S solution. Regarding the use of LLA, which also requires a starting value specification, we again examined the cases for which LLA resulted in lower objective function evaluations. For these cases, all MIST solutions were LLA solutions, and all LLA solutions were MIST solutions with the exception of one. Hence, both methods find valid, if often different, solutions, a behavior that we again attribute to the differences in paths taken towards a solution.

4.3. Effectiveness of convergence acceleration

We explored the effectiveness of SQUAREM^2, defined in Section 3.3, when applied to several simulated datasets taken from the previous two simulation studies. Table 6 indicates the relative reduction in elapsed time ('RRT') and numbers of MM updates, i.e., invocations of mapping $M(\cdot)$, required for the original and SQUAREM^2-accelerated algorithms to converge for five randomly chosen simulation datasets across the five penalty functions. The SQUAREM^2 algorithm converged without difficulty in these cases and required substantially fewer MM updates than the original algorithm; the percent reduction in time was as high as 96%. We remark here that the regularity conditions imposed in Roland and Varadhan [53] and Varadhan and Roland [67], particularly smoothness conditions, are not satisfied in this particular class of examples. Hence, while the simulation results are certainly very promising, the question of convergence (and its associated rate) of SQUAREM^2 in this class of problems continues to remain an interesting open problem.

TABLE 6

Acceleration from SQUAREM² applied to simplified MIST algorithm for five randomly selected simulation datasets. The relative reduction in elapsed time is given by 'RRT', while the number of MM updates are given for the original MIST implementation and SQUAREM² implementation in '# orig' and '# sm² ', respectively. Parameter Θ in the top portion of the table collects the dimension and noise information for the linear model examples, i.e., Θ = (p, σ)

	LAS			ALAS			EN			AEN			SCAD		
Dataset	RRT	#orig	#sm²	RRT	#orig	#sm²	RRT	#orig	#sm²	RRT	#orig	#sm²	RRT	#orig	#sm²
LM															
Θ = (35, 1)															
62	0.67	260	62	0.81	169	44	0.63	46	26	0.82	42	23	0.91	485	68
71	0.76	221	59	0.75	163	41	0.67	49	29	0.62	44	29	0.83	302	65
86	0.67	271	68	0.70	149	44	0.67	51	29	0.75	43	26	0.93	987	104
95	0.86	317	74	0.88	187	41	0.92	49	29	0.73	46	26	0.90	500	71
88	0.88	330	68	0.87	162	41	0.78	51	29	0.77	45	26	0.90	528	77
Θ = (81, 3)															
62	0.90	2059	242	0.89	589	92	0.65	68	35	0.75	64	29	0.88	594	101
71	0.93	1426	164	0.93	838	83	0.76	77	32	0.70	71	32	0.94	2608	215
86	0.90	1351	212	0.92	956	98	0.59	77	38	0.79	69	32	0.92	1038	110
95	0.93	1500	167	0.86	367	71	0.67	72	35	0.74	68	29	0.90	663	92
88	0.92	1547	185	0.90	716	101	0.60	70	32	0.68	66	32	0.92	1798	203
GLM															
$q = 25$															
62	0.93	4928	431	0.96	6227	272	0.89	3201	359	0.93	3316	236	0.95	22044	1442
71	0.92	4195	416	0.95	5045	239	0.90	2796	281	0.94	2843	170	0.95	16225	1052
86	0.92	4488	470	0.95	5449	242	0.92	2971	257	0.93	3044	206	0.95	20133	1193
95	0.93	4553	374	0.94	5419	341	0.92	3059	269	0.95	3096	152	0.95	15250	1064
88	0.92	5212	527	0.95	6850	371	0.91	3237	314	0.94	3393	203	0.96	26477	1367
$q = 75$															
62	0.88	4334	674	0.91	3573	377	0.85	3055	575	0.90	2435	293	0.95	88994	5687
71	0.91	3805	446	0.92	3046	281	0.85	2761	536	0.89	2194	281	0.94	82615	5588
86	0.87	3615	602	0.91	2900	329	0.87	2653	434	0.92	2110	185	0.93	42652	3686
95	0.89	3870	554	0.90	3121	380	0.90	2820	338	0.89	2264	314	0.94	40002	3095
88	0.88	4177	641	0.94	3395	251	0.87	2972	482	0.91	2415	242	0.94	77484	5885

5. Example: Genes associated with lymphoma patient survival

Diffuse large-B-cell lymphoma (DLBCL) is an aggressive type of non-Hodgkins lymphoma and is one of the most common forms of lymphoma occurring in adults. Rosenwald *et al.* [54] utilized Lymphochip DNA microarrays, specialized to include genes known to be preferentially expressed within the germinal centers of lymphoid organs, to collect and analyze gene expression data from 240 biopsy samples of DLBCL tumors. For each subject, 7399 gene expression measurements were obtained. The expression profiles along with corresponding patient information can be downloaded from their supplemental website *http://llmpp.nih.gov/DLBCL/*. Since the original profiles had some missing expression measurements, we used the dataset subsequently analyzed by Li and Gui [40] which estimated the missing values using a nearest neighbor approach. During the time of followup, 138 patient deaths were observed with median death time of 2.8 years.

Rosenwald *et al.* [54] used hierarchical clustering to group the genes into four gene-expression signatures: Proliferation (PS), which includes cell-cycle control and checkpoint genes, and DNA synthesis and replication genes; Major Histocompatibility Complex ClassII (MHC), which includes genes involved in antigen presentation; Lymph Node (LNS), which includes genes encoding for known markers of monocytes, macrophages, and natural killer cells; and Germinal Cen-

ter B (GCB), which includes genes that are characteristic of germinal center B cells; see Alizadeh *et al.* [1] for more information on gene signatures. Based on the gene clusters, they built a Cox proportional hazards model [14, 15] to predict survival outcomes in the DLBCL patients. Subsequently, this dataset has been analyzed numerous times, typically to evaluate methods related to subgroup identification and/or survival prediction [e.g., 4, 20, 27, 28, 40, 41, 63].

Here, we instead focus on the performance of two different penalties, namely SCAD and MCP, with regard to the identification of genes associated with DLBCL survival. The simulation results of Zhang [71, 72] suggest that MCP has superior selective accuracy over the SCAD penalty, at least for the case of a linear model. There, selection accuracy was measured as the proportion of simulation replications with correct classification of both the zero and non-zero coefficients, with MCP outperforming SCAD in all simulation settings.

To illustrate the utility and flexibility of the MIST algorithm, we reanalyzed the DLBCL data, fitting a penalized Cox regression model respectively using SCAD and MCP penalty functions, and running these procedures in combination with the Iterative Sure Independence Screening procedure [ISIS, 23] in order to ensure that the dimension of the parameter space was maintained at a manageable level. For SCAD, we considered both the 1S and LLA estimators. The existing optimization functions provided in the *SIS* package for the ISIS procedure were used for the 1S estimator, whereas relevant modifications to the ISIS code were made in order to accommodate the fully iterative LLA and MCP estimators. Optimization at each step of the ISIS algorithm in the case of the MCP penalty utilized the MIST algorithm, as we are aware of no other algorithm capable of fitting the Cox regression model subject to MCP penalization. The default settings in the *SIS* package were used to determine the maximum number of predictors ($[\frac{n}{4\log n}] = 10$) and to define the additional ISIS parameters (e.g., use of the unpenalized MLE as a starting value, ranking method, tuning parameter selection) for all three analyses (1S-SCAD, LLA-SCAD, MIST-MCP). The parameter a was set to 3.7 for all analyses; hence, only the selection of λ required any tuning.

Table 7 displays the 11 genes identified by at least one of the three analyses. The x's in a given column indicate the genes with non-zero coefficients resulting from the corresponding penalization. The final column provides references for genes previously linked to DLBCL in the literature. Genes belonging to the original Rosenwald *et al.* [54] gene expression signatures are indicated with parenthetical initials. Note that the references provided are not meant to be an exhaustive list, but instead intend to demonstrate the relevance of certain genes and/or their altered expression levels in DLBCL survival.

Interestingly, the LLA-SCAD and MIST-MCP penalizations selected the same subset of genes, having a nearly complete overlap with those selected from the 1S-SCAD penalization. The number of genes selected in each case is 10, the maximum specified by ISIS; 9 of these were shared across the three penalizations. According to NCBI Entrez Gene search (http://www.ncbi.nlm.nih.gov/), many of these genes are biologically relevant. For example, CDK7 codes for a protein

TABLE 7
Genes associated with DLBCL survival with SCAD (one-step=1S and LLA) and MCP penalizations, sorted by the gene order in the original data set. ID refers to the unique Lymphochip identification number. The x's in a given column indicate the genes identified by the corresponding penalization

ID	Name (Symbol)	SCAD 1S	SCAD LLA	MCP	References
27774	cyclin-dependent kinase 7 (CDK7)	x	x	x	[54] (PS), [43], [7, 8]
31242	acidic 82 kDa protein mRNA (DNTTIP2)	x	x	x	[7, 8]
31981	septin 1 (SEPT1)	x	x	x	[54] (PS), [41], [60] [58], [73], [4], [7, 8]
29652	BUB3 budding uninhibited by benzimidazoles 3 (BUB3)	x	x	x	[54] (PS)
27731	major histocompatibility complex, class II, DR alpha (HLA-DRA)	x	x	x	[54] (MHC), [41] [27, 28], [61], [8]
24376	ESTs, Weakly similar to A47224 thyroxine-binding globulin precursor	x	x	x	[54] (GCB), [3], [41] [27, 28], [4], [61] [7, 8]
22162	delta sleep inducing peptide, immunoreactor (TSC22D3)		x	x	[44]
23862	(AI568329) ESTs	x	x	x	
24271	integrin, alpha L (ITGAL)	x			[44]
33358	(AA830781)	x	x	x	[41], [8]
32679	KIAA0084 protein (RFTN1)	x	x	x	[28], [58], [73], [4] [7, 8]

that regulates cell cycle progression and is represented in the Proliferation Signature, although reported under a different Lymphochip ID as this gene was spotted multiple times on the array. Also members of the Proliferation Signature are SEPT1, coding for a protein involved in cytokinesis, and BUB3, coding for a mitotic checkpoint protein. DNTTIP2 regulates transcriptional activity of DNTT, a gene for a protein expressed in a restricted population of normal and malignant pre-B and pre-T lymphocytes during early differentiation. HLA-DRA, a member of the MHC Signature, plays a central role in the immune system and is expressed in antigen presenting cells, such as B lymphocytes, dendritic cells, macrophages. From the GCB Signature, the ESTs weakly similar to thyroxine-binding globulin precursor is highly cited. Additionally, RFTN1 plays a pivotal role in regulating B-cell antigen receptor-mediated signaling [56].

The gene AI568329, selected by all methods, is not described in the original dataset and its function is unknown. Similarly, although cited at least twice, a description for AA830781 is also unavailable. Both of these genes may be related to lymphoma or risk of death from lymphoma, as it is possible that these genes (and potentially others) were selected because of coexpression or correlation with other relevant genes. Interestingly, the two genes not commonly identified across the three penalizations were both cited in Martinez-Climent *et al.* [44]. They found altered gene expression of TSC22D3 and ITGAL (both involved in a variety of immune phenomena) in one case who initially presented with follicle center lymphoma and subsequently transformed to DLBCL.

The results of this analysis demonstrate equivalence in selection performance between MCP and LLA-SCAD for the case of Cox proportional hazards model. Increasing the maximum number of predictors to 21 again resulted in equivalent selection performance between MCP and LLA-SCAD, with 21 predictors ultimately selected (results not shown). The 1S estimator also resulted in the selection of 21 predictors, but with increased dissimilarity between MCP/LLA-SCAD and 1S: only 13 of the 21 genes were selected by all three methods. It should be noted that Zhang [71, 72] did not use any form iterative variable selection (e.g., ISIS) in his comparisons between SCAD and MCP for the case of the linear model; in addition, Zhang [71, 72] fixed values for both penalty parameters in his simulations and also did not use $a = 3.7$. The use of ISIS, the methodology used for selecting λ, and the use of $a = 3.7$ [e.g., 23] in both the MCP and SCAD penalties may all play a role in the results summarized above.

6. Discussion

This paper proposed a versatile and general algorithm capable of dealing with a wide variety of nonsmoothly penalized objective functions, including but not limited to all presently popular combinations of goodness-of-fit and penalty functions. In particular, the MIST algorithm utilizes a judicious choice of majorization to generate a MM algorithm that applies soft-thresholding componentwise and which, in certain settings, allows one to minimize the majorizing function in a single iteration. We established a suitable convergence theory, as well as new results on the convergence of rather general MM algorithms. In the case of penalized least squares, our results are complementary to convergence results obtained for coordinate descent algorithms designed for use with the lasso penalty [cf. 45, 65, 69]. In general, while the minimizers obtained at each step of the MIST algorithm are not necessarily coordinatewise minima of the desired objective function, the MIST algorithm continues to drive the objective function steadily downhill, converging to a local minimizer. We further demonstrated the remarkable effectiveness of the simple SQUAREM^2 acceleration procedure in these problems as tool for accelerating the slow but steady convergence of the proposed class of algorithms. Beyond specification of the penalty parameter(s) $\boldsymbol{\lambda}$, virtually no effort was expended in tuning or otherwise specializing the MIST algorithm for solving a given problem. At the expense of greater analytical work, the rate of convergence for the standard MIST algorithm can itself be improved.

The simulation results of this paper highlight the fact that nonconvex penalties tend to endow the corresponding objective function with multiple local minima. The resulting sensitivity of computational algorithms to the choice of starting value, while known, has not been especially emphasized in the literature on penalized estimation. In this regard, the computationally attractive one-step method of Zou and Li [77] provides a useful choice of starting value for fully iterative SCAD-based algorithms. In addition to being unique under mild regularity conditions, it is also easily generalized to other nonconvex penalties, such as MCP, and yields estimators with attractive asymptotic properties. Unfortunately, its utility for identifying starting values is limited to settings where

$N > p$ because the justification for the 1S estimator relies heavily on the use of the unpenalized MLE as a starting value.

SparseNet [45] provides an interesting addition to the class of methods able to deal with non-convex penalty functions in the case of penalized least squares. The implementation of this methodology using the MCP penalty [71, 72] appears to hold particular promise as a tool for variable selection. As indicated earlier, the core optimization procedure used within SparseNet is a form of iterated thresholding and is developed with a particular focus on the linear model. It would be interesting to explore the possibility of extending SparseNet to the class of generalized linear models and related problems (e.g., Cox regression), with one obvious approach being to replace the coordinate descent algorithm with the MIST algorithm.

The simulated examples in this paper only consider settings with $N > p$, in part to ensure that the goodness-of-fit function $g(\boldsymbol{\beta})$ remains strictly convex. While the MIST algorithm has not yet been extensively tested in settings where $N \ll p$, preliminary results show that the algorithm continues to find reasonable solutions when given a reasonable starting value, but tends to converge at a slower rate in comparison with $N > p$. As suggested by Table 6, the SQUAREM^2 acceleration procedure can produce dramatic gains even as N gets close to p; the problem of tuning the algorithm, the development of acceleration procedures and the problem of selecting suitable starting values in problems with multiple local minima, particularly in settings where $p > N$ but the number of "important" predictors $p_0 \ll N$, are left for future work. Two important and unresolved challenges in such problems include rigorously justifiable methods for determining good starting values and penalty parameter(s).

Appendix A

This appendix is divided into several sections. Section A.1 reviews and extends the convergence theory for the EM algorithm established in Wu [68]; the extension utilizes results of Meyer [47] to establish stronger convergence results for general MM algorithms. Section A.2 contains the proof of Theorem 2.1 and makes direct use of these results. Finally, Sections A.3 and A.4 respectively contain the proofs of Theorems 3.2 and 3.4, establishing the convergence of iterated soft thresholding when used to minimize (10) and convergence of the proposed class of MIST algorithms in the case of the generalized linear model.

A.1. Local convergence of MM algorithms in nonsmooth problems

Using convergence theory for algorithms derived from point-to-set maps developed by Zangwill [70], Wu [68] established some general convergence results for the EM algorithm under a range of conditions. In what follows, the key convergence result of Zangwill [70] is restated; this result, given in Theorem A.1 and adapted from Wu [68], is stated in a form convenient for use with the MM algorithm and provides for a very general (and comparatively weak) form of

convergence. We then draw on stronger convergence results due to Meyer [47] in order to establish a more useful convergence theory for MM algorithms designed to minimize nondifferentiable objective functions; this result is stated in Theorem A.3. Finally, we provide a set of sufficient regularity conditions that ensure the validity of the conditions of both theorems in a wide class of statistical estimation problems.

Let $\xi(\boldsymbol{\beta})$ be the real-valued function to be minimized, where $\boldsymbol{\beta} \in \mathcal{B}$. While the focus of this paper is on $\mathcal{B} = \mathbb{R}^p$, the development below assumes only that $\mathcal{B}$ is some (possibly proper) subset of $\mathbb{R}^p$ and that $\mathcal{B}_0 \subset \mathcal{B}$ is a compact subset. Let $M : \mathcal{B} \to \mathcal{B}$ denote the map (1), where $\xi^{[\mathrm{S}]}(\cdot, \cdot)$ is any function that majorizes $\xi(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \mathcal{B}$. In general, $M(\cdot)$ is a point-to-set map, and therefore a set. We say that $\bar{\boldsymbol{\beta}}$ is a generalized fixed point of $M(\cdot)$ if $\bar{\boldsymbol{\beta}} \in M(\bar{\boldsymbol{\beta}})$; we say that $\bar{\boldsymbol{\beta}}$ is a fixed point of $M(\cdot)$ if $M(\bar{\boldsymbol{\beta}}) = \{\bar{\boldsymbol{\beta}}\}$ (i.e., a singleton). Theorem A.1 below states Theorem A of Zangwill [70] for the case of the MM algorithm.

Theorem A.1. *Suppose $\xi(\boldsymbol{\beta})$ is a continuous, real-valued function of $\boldsymbol{\beta} \in \mathcal{B}$ that is uniformly bounded below. Assume $\boldsymbol{\beta}^{(0)} \in \mathcal{B}$ is a bounded vector and that $\xi(\boldsymbol{\beta}^{(0)}) < \infty$. Let the sequence $\{\boldsymbol{\beta}^{(n)}, n \geq 0\}$ be generated as follows: $\boldsymbol{\beta}^{(n+1)} \in M(\boldsymbol{\beta}^{(n)})$, where $M(\cdot)$ is the point-to-set map (1). Let $\mathcal{S} \subset \mathcal{B}$ denote a specified non-empty solution set. Suppose that*

Z1. Each $\boldsymbol{\beta}^{(n)} \in \mathcal{B}_0$, $n \geq 0$;
Z2. $M(\cdot)$ is closed and non-empty for $\boldsymbol{\beta} \in \mathcal{S}^c$;
Z3. The following two conditions hold:

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

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

One may then draw the following conclusions:

M1. The sequence $\{\boldsymbol{\beta}^{(n)}, n \geq 0\}$ has at least one limit point in $\mathcal{S}$; in addition, the set of all such points, say $\mathcal{S}_0$, satisfies $\mathcal{S}_0 \subseteq \mathcal{S}$.
M2. Each limit point $\bar{\boldsymbol{\beta}} \in \mathcal{S}_0$ satisfies $\lim_{n\to\infty} \xi(\boldsymbol{\beta}^{(n)}) = \xi(\bar{\boldsymbol{\beta}})$.
M3. Each limit point $\bar{\boldsymbol{\beta}} \in \mathcal{S}_0$ is a generalized fixed point of $M(\cdot)$.

Remark A.2. *Assumptions [Z1]-[Z3] are imposed in Wu [68]. The assumption [Z1] implies that $\{\boldsymbol{\beta}^{(n)}, n \geq 0\}$ is a bounded sequence, ensuring the existence of at least one limit point. Further comments on [Z2] will be made below, as it is possible to impose reasonable sufficient conditions that ensure this condition. The assumption [Z3] enforces the descent property at each update, as would be expected in any EM, GEM or MM algorithm. An equivalent formulation of [Z3] follows [e.g., 47, p. 114]:*

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

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

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

Conclusion [M1] means the limit of any convergent subsequence $\boldsymbol{\beta}^{(n_j)}$ lies in $\mathcal{S}$; hence, the above theorem guarantees convergence of subsequences, but does not ensure global convergence of the iteration sequence. Subsequential convergence permits, for example, oscillatory behavior in the limit sequence. Meyer [47, 48] offers several refinements of Theorem A.1, strengthening the statements of convergence. His results, adapted for the MM algorithm, follow below; in particular, see Theorem 3.1, Corollary 3.2, and Theorems 3.5 and 3.6 of Meyer [47].

Theorem A.3. *Let the conditions of Theorem A.1 hold. Define the conditions:*

- *Z4. For each $\boldsymbol{\alpha} \in \mathcal{B}$ and any $\boldsymbol{\beta} \in M(\boldsymbol{\alpha})$, we have $\xi(\boldsymbol{\beta}) < \xi(\boldsymbol{\alpha})$ whenever $M(\boldsymbol{\alpha}) \neq \{\boldsymbol{\alpha}\}$ (i.e., a strict decrease in the objective function occurs at any point $\boldsymbol{\alpha}$ that is not a fixed point);*
- *Z5. there exists an isolated point $\bar{\boldsymbol{\beta}}^*$ such that $M(\bar{\boldsymbol{\beta}}^*) = \{\bar{\boldsymbol{\beta}}^*\}$ (i.e., a true fixed point).*

Suppose [Z1]-[Z4] hold. Then, in addition to results [M1]-[M3] of Theorem A.1, the following conclusions hold:

- *M4. Each limit point $\bar{\boldsymbol{\beta}} \in \mathcal{S}_0$ satisfies $M(\bar{\boldsymbol{\beta}}) = \{\bar{\boldsymbol{\beta}}\}$, and is therefore a fixed point of $M(\cdot)$;*
- *M5. $\lim_{n\to\infty} \|\boldsymbol{\beta}^{(n)} - \boldsymbol{\beta}^{(n+1)}\| = 0$, in which case one either has (i) convergence of $\boldsymbol{\beta}^{(n)}$ to a limit; (ii) the set of limit points of $\boldsymbol{\beta}^{(n)}$ forms a continuum, hence $\boldsymbol{\beta}^{(n)}$ fails to converge;*
- *M6. If the number of fixed points having any given value of $\xi(\cdot)$ is finite, then $\{\boldsymbol{\beta}^{(n)}, n \geq 0\}$ converges to one of these fixed points;*
- *M7. If the sequence $\{\boldsymbol{\beta}^{(n)}, n \geq 0\}$ has an isolated fixed point $\bar{\boldsymbol{\beta}}$, then $\boldsymbol{\beta}^{(n)} \to \bar{\boldsymbol{\beta}}$. If $\bar{\boldsymbol{\beta}}$ is also an isolated local minimum of $\xi(\cdot)$ on $\mathcal{B}_0$, then there exists an open neighborhood $\mathcal{B}_\epsilon \subseteq \mathcal{B}_0$ of $\bar{\boldsymbol{\beta}}$ such that $\boldsymbol{\beta}^{(n)} \to \bar{\boldsymbol{\beta}}$ if $\boldsymbol{\beta}^{(0)} \in \mathcal{B}_\epsilon$.*

Suppose instead that [Z1-Z3] and [Z5] hold. Then, in addition to results [M1]-[M3] of Theorem A.1, the following conclusion can be drawn:

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

Remark A.4. *Assumption [Z4] strengthens [Z3] by imposing the condition that the iteration scheme is strictly monotonic; as such, all generalized fixed points of $M(\cdot)$ are also fixed points, a situation that permits stronger statements of convergence results. Assumption [Z5] imposes the somewhat weaker assumption that there exists at least one isolated fixed point of the iteration sequence; similarly to [M7], [M8] implies that the iteration converges to this point. Two further consequences of these results are (i) one may take $\mathcal{S}$ to be the set of fixed points of $M(\cdot)$; and, (ii) all solutions to the minimization problem $\min_{\boldsymbol{\beta}\in\mathcal{B}} \xi(\boldsymbol{\beta})$ are in fact fixed points of $M(\cdot)$, hence contained within $\mathcal{S}$ [47, pp. 110-11].*

Conclusions [M1]-[M7] essentially mirror those in Vaida [66, Theorems 1-3], who obtains strong convergence results for the EM and MM algorithms un-

der continuous differentiability assumptions on the objective and majorization functions and the additional condition that $\xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ has a unique global minimizer in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha} \in \mathcal{S}$, where $\mathcal{S}$ is the (assumed finite) set of stationary points of $\xi(\boldsymbol{\beta})$. This uniqueness condition, reflected in [Z4], provides a verifiable convergence condition that is often satisfied in statistical applications.

Sufficient conditions that ensure [Z1]-[Z4], but weaker than conditions imposed in Vaida [66], are now provided. In particular, suppose that the objective function, its surrogate and the mapping $M(\cdot)$ satisfy the following regularity conditions:

R1. $\xi(\boldsymbol{\beta})$ is locally Lipschitz continuous for $\boldsymbol{\beta} \in \mathcal{B}$ and there exists at least one $\mathbf{b}_0 \in \mathcal{B}$ such that $L(\xi(\mathbf{b}_0)) = \{\mathbf{b} \in \mathcal{B} : \xi(\mathbf{b}) \leq \xi(\mathbf{b}_0)\}$ is compact. Assume that the set of stationary points $\mathcal{S}$ of $\xi(\boldsymbol{\beta})$ is a finite set, where the notion of a stationary point is defined as in Clarke [12].
R2. $\xi(\boldsymbol{\beta}) = \xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\beta})$ for each $\boldsymbol{\beta} \in \mathcal{B}$.
R3. $\xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\alpha}) > \xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\beta})$ for $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$, $\boldsymbol{\beta}, \boldsymbol{\alpha} \in \mathcal{B}$.
R4. $\xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is continuous for $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{B} \times \mathcal{B}$ and locally Lipschitz in $\boldsymbol{\beta}$ for $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$.
R5. $M(\boldsymbol{\beta})$ is a singleton set consisting of one bounded vector for each $\boldsymbol{\beta} \in \mathcal{B}$.

The above conditions do not imply that the objective function $\xi(\boldsymbol{\beta})$ is differentiable everywhere. Condition [R1] does imply that $\xi(\boldsymbol{\beta})$ is bounded below on $\mathcal{B}$, that $\nabla\xi(\boldsymbol{\beta})$ exists for almost all $\boldsymbol{\beta}$, and that the set of global minimizers of $\xi(\boldsymbol{\beta})$ on $\mathcal{B}$ is non-empty and bounded. Conditions [R2] and [R3] imply that $\xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ strictly majorizes $\xi(\boldsymbol{\beta})$ and, in addition,

$$\xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \xi(\boldsymbol{\beta}) + \psi(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{21}$$

where $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) := \xi^{[S]}(\boldsymbol{\beta}, \boldsymbol{\alpha}) - \xi(\boldsymbol{\beta})$ satisfies $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 0$ for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ and $\psi(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$. Assumptions [R4] and [R5] imply that the map $M(\boldsymbol{\beta})$ is continuous, hence bounded on compact sets [52, Proposition 3.2]. Conditions [R1], [R4], and [R5] further imply that (21) is bounded below for $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{B} \times \mathcal{B}$ and that $\psi(\bar{\boldsymbol{\beta}}, \boldsymbol{\alpha})$ is uniquely minimized at $\boldsymbol{\alpha} = \bar{\boldsymbol{\beta}}$ for any fixed point $\bar{\boldsymbol{\beta}}$.

Suppose conditions [R1]-[R5] hold. As commented earlier, [R4] and [R5] imply that $M(\boldsymbol{\beta})$ is a continuous point-to-point map; hence, $M(\cdot)$ is closed [e.g., 42, pp. 203-204], establishing [Z2]. Propositions A.6 and A.7, given below and proved under [R1]-[R5] in Schifano [57], establish [Z1], [Z3] and [Z4]. An important consequence of the sufficient conditions [R1]-[R5] is that the set of fixed points for the mapping $M(\cdot)$ also coincides with the set of stationary points for $\xi(\cdot)$; see Proposition A.8.

Remark A.5. *Condition [R1] refers to Clarke [12] for the definition of a stationary point; see also Theorems 2.1 and 3.4. Because $\xi(\boldsymbol{\beta})$ is assumed to be locally Lipschitz continuous, a point $\boldsymbol{\beta}^*$ is a stationary point in the sense of Clarke if $\mathbf{0} \in \partial\xi(\boldsymbol{\beta}^*)$, where $\partial\xi(\boldsymbol{\beta})$ denotes the Clarke subdifferential of $\xi(\boldsymbol{\beta})$ [12, Proposition 2.3.2]. The condition that $\mathbf{0} \in \partial\xi(\boldsymbol{\beta}^*)$ is a necessary but not sufficient condition for $\nabla\xi(\boldsymbol{\beta}^*) = \mathbf{0}$. That is: if $\nabla\xi(\boldsymbol{\beta}^*) = \mathbf{0}$, then $\partial\xi(\boldsymbol{\beta}^*) = \{\nabla\xi(\boldsymbol{\beta}^*)\} = \{\mathbf{0}\}$;*

however, $\nabla\xi(\boldsymbol{\beta}^*)$ need not exist in order for $\mathbf{0} \in \partial\xi(\boldsymbol{\beta}^*)$. In general, the assumption that $\mathcal{S}$ is finite does not mean that the gradient exists at any of these points; in view of Proposition A.8, conditions [R1]-[R5] also do not imply an equivalence between the existence of an isolated fixed point of $M(\cdot)$ and differentiability of $M(\cdot)$ at that point.

Proposition A.6. *Let $n \geq 0$ be given and suppose $\boldsymbol{\beta}^{(n)} \in \mathcal{B}$ is a bounded vector. Then, $\boldsymbol{\beta}^{(n+1)} = M(\boldsymbol{\beta}^{(n)}) \in \mathcal{B}$ is bounded, and is unique. In addition, for $n \geq 0$,*

$$\xi^{[S]}(\boldsymbol{\beta}^{(n+1)}, \boldsymbol{\beta}^{(n)}) \leq \xi^{[S]}(\boldsymbol{\beta}^{(n)}, \boldsymbol{\beta}^{(n)}) < \infty \tag{22}$$

and

$$\xi(\boldsymbol{\beta}^{(n+1)}) - \xi(\boldsymbol{\beta}^{(n)}) \leq -\psi(\boldsymbol{\beta}^{(n+1)}, \boldsymbol{\beta}^{(n)}) \leq 0, \tag{23}$$

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

Proposition A.7. *Let $\boldsymbol{\beta}^{(0)} \in \mathcal{B}$ be a bounded vector and set $\xi^{(n)} = \xi(\boldsymbol{\beta}^{(n)})$ for $n \geq 0$. Then, $\{\xi^{(n)}, n \geq 0\}$ is a bounded, monotone decreasing sequence and $\boldsymbol{\beta}^{(n)} \in L(\xi^{(0)}) \subset \mathcal{B}$ for every $n \geq 0$, where $L(\xi^{(0)})$ is a compact set.*

Proposition A.8. *Under [R1]-[R5], the set of fixed points for the mapping $M(\cdot)$ coincides with the set of stationary points for $\xi(\cdot)$, where the notion of a stationary point is defined as in Clarke [12].*

Proof of Proposition A.8:

Let $\bar{\boldsymbol{\beta}}$ be a fixed point of $M(\cdot)$. Since $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is locally Lipschitz continuous for $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$ for each bounded $\boldsymbol{\alpha}$, the relation $\bar{\boldsymbol{\beta}} = M(\bar{\boldsymbol{\beta}})$ is equivalent to

$$\mathbf{0} \in \partial\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}},$$

where the right-hand side denotes the Clarke subdifferential of $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})$ with respect to $\boldsymbol{\beta}$, evaluated at $\boldsymbol{\beta} = \bar{\boldsymbol{\beta}}$. Using Proposition 2.3.3 of Clarke [12, p. 38],

$$\partial\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} \subset \partial\xi(\bar{\boldsymbol{\beta}}) \ \oplus \ \partial\psi(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}},$$

where the right-hand side denotes the set consisting of all elements $a + b$, where $a \in \partial\xi(\bar{\boldsymbol{\beta}})$ and $b \in \partial\psi(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}}$. It follows that $\bar{\boldsymbol{\beta}}$ is a stationary point of $\xi(\boldsymbol{\beta})$ if $\partial\psi(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} = \{\mathbf{0}\}$, since in this case we have $\partial\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} = \partial\xi(\bar{\boldsymbol{\beta}})$ and hence that $\mathbf{0} \in \partial\xi(\bar{\boldsymbol{\beta}})$.

To establish that $\partial\psi(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} = \{\mathbf{0}\}$, we recall that $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is locally Lipschitz continuous in $\boldsymbol{\beta}$ for $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$ and additionally satisfies $\psi(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$ and $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 0$ for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ As stated earlier, conditions [R1], [R4], and [R5] further imply that (21) is bounded below for $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathcal{B} \times \mathcal{B}$ and that $\psi(\bar{\boldsymbol{\beta}}, \boldsymbol{\alpha})$ is uniquely minimized at $\boldsymbol{\alpha} = \bar{\boldsymbol{\beta}}$ for any fixed point $\bar{\boldsymbol{\beta}}$. Hence, $\partial\psi(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}})|_{\boldsymbol{\beta}=\bar{\boldsymbol{\beta}}} = \{\mathbf{0}\}$ as desired. □

A.2. Proof of Theorem 2.1

The assumptions stated in the theorem immediately yield that $\xi(\boldsymbol{\beta})$ is locally Lipschitz continuous for each bounded $\lambda > 0$, hence (i) is satisfied; in addition, the stated assumptions imply $\xi(\boldsymbol{\beta})$ is also coercive, hence attains a global minimum interior to $\mathbb{R}^p$.

To show (ii), we first write

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

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

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

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

Thus, the claimed equality between (3) and (4) holds in this case.

Now, suppose that $\tilde{p}(r; \boldsymbol{\theta})$ is nonlinear in r. Under (P1), we claim that (4) strictly majorizes $p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ provided the derivative of the penalty $\tilde{p}'(\cdot, \boldsymbol{\lambda}_j)$ is strictly positive. To see this, observe that concavity (e.g., see (6)) implies the inequality

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

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

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

and (ii) is established.

Define $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ so that $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha}) \equiv \xi(\boldsymbol{\beta}) + \psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$. In order to establish the majorization property specified in (iii), we begin by noting that our assumptions on $g(\boldsymbol{\beta})$, $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$, and $\tilde{p}(\cdot; \boldsymbol{\theta})$ imply that $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are both continuous in $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Our assumptions further imply that $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) \geq 0$; if at least one of $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ or $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is strictly positive for $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$, then $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) > 0$ for $\boldsymbol{\alpha} \neq \boldsymbol{\beta}$ and $\psi(\boldsymbol{\beta}, \boldsymbol{\beta}) = 0$. Therefore, the objective function $\xi(\boldsymbol{\beta})$ is strictly majorized by $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \xi(\boldsymbol{\beta}) + \psi(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

In order to establish the convergence of the corresponding MM algorithm in (iii), it suffices to prove that the assumptions of the theorem and consequent assertions established thus far are sufficient to ensure that Conditions

[R1]-[R5] of Appendix A.1 are met, in which case Theorem A.3 applies directly. The result (i), combined with the fact that $\xi(\boldsymbol{\beta})$ attains a global minimum and the assumption that its corresponding set of stationary points is also finite, immediately establishes [R1]; as proved above, [R2] and [R3] also hold. If $\psi(\boldsymbol{\beta}, \boldsymbol{\alpha}) = h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is continuous in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and locally Lipschitz continuous in $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$, then (i) implies that [R4] also holds. By assumption, $h(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is continuous in $\boldsymbol{\alpha}$ and continuously differentiable in $\boldsymbol{\beta}$, hence locally Lipschitz in $\boldsymbol{\beta}$. Continuity of $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ in both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is also immediate. Hence, [R4] holds provided that $q(\boldsymbol{\beta}, \boldsymbol{\alpha}; \boldsymbol{\lambda}) - p(\boldsymbol{\beta}; \boldsymbol{\lambda})$ is locally Lipschitz continuous in $\boldsymbol{\beta}$ near $\boldsymbol{\alpha}$. To see that this is the case, we note that (24) is a linear combination of functions in β_j of the form $\tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j| - \tilde{p}(|\beta_j|; \boldsymbol{\lambda}_j)$, where $|\cdot|$ and $-\tilde{p}(\cdot; \boldsymbol{\lambda})$ are both convex, hence locally Lipschitz. Since both the sum and composition of two locally Lipschitz functions are locally Lipschitz, the result now follows. Finally, [R5] is ensured by [R1]-[R4] and the condition in (iii) that $\xi^{[\mathrm{S}]}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is uniquely minimized in $\boldsymbol{\beta}$ for each $\boldsymbol{\alpha}$.

A.3. Proof of Theorem 3.2

Under the stated conditions and for any bounded $\boldsymbol{\alpha}$, $m(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + h(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \lambda\varepsilon\|\boldsymbol{\beta}\|^2$ is strictly convex with a Lipschitz continuous derivative of order $L^{-1} > 0$; in addition, $\sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|$ is also convex in $\boldsymbol{\beta}$. Hence, for each bounded $\boldsymbol{\alpha}$ there exists a unique solution $\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\boldsymbol{\alpha})$ when minimizing (10).

In the notation of Combettes and Wajs [13], we may identify the Hilbert space $\mathcal{H}$ with $\mathbb{R}^p$, $f_2(\boldsymbol{\beta})$ with $m(\boldsymbol{\beta})$ and $f_1(\boldsymbol{\beta})$ with $\sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|$. The assumptions of the theorem ensure that the regularity conditions of Proposition 3.1 and Theorem 3.4 of Combettes and Wajs [13] are met. In particular, because $m(\boldsymbol{\beta})$ is strictly convex, Proposition 3.1 guarantees the existence of a unique solution to the desired optimization problem as well as provides the relevant fixed point mapping; Theorem 3.4 establishes the convergence of the corresponding iterative scheme to this unique solution.

Proposition 3.1 and Theorem 3.4 of Combettes and Wajs [13] each rely on the gradient of $f_2(\boldsymbol{\beta})$ and the "proximity operator" of $f_1(\boldsymbol{\beta})$. In the present setting, Example 2.20 in Combettes and Wajs [13] shows that the proximity operator for $\sum_{j=1}^{p} \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\beta_j|$ is exactly $S(\cdot; \boldsymbol{\tau})$; see Step 2. The algorithm summarized in the statement of the theorem is now just a specific instance of that described in the Theorem 3.4 with (in their notation) $a_n = b_n = 0$ and $\lambda_n = 1$ for every n.

Hale, Yin and Zhang [29, Theorem 4.5] undertake a detailed study of the proposed algorithm for the special case of a convex, differentiable $f_2(\boldsymbol{\beta})$ and $f_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$. In this case, they prove that the algorithm converges in a finite number of iterations. A minor extension of their arguments may be used to establish the same result for the algorithm described here.

A.4. Proof of Theorem 3.4

First, we note that assumptions of this theorem are sufficient to ensure that $\xi(\widetilde{\boldsymbol{\beta}})$ is locally Lipschitz continuous on $\mathbb{R}^{p+1}$. To establish (1.), note that the choice of $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ in (13) with appropriately chosen ϖ guarantees majorization of $-\ell(\widetilde{\boldsymbol{\beta}})$ since our assumptions imply that $\nabla^2(-\ell(\widetilde{\boldsymbol{\beta}}))$ can be suitably bounded on $\mathbb{R}^{p+1}$ [e.g., 38, Chapter 6]. As shown earlier, penalties of the form (3) satisfying assumption (P1) can also be linearly majorized. Hence, (14) majorizes $\xi(\widetilde{\boldsymbol{\beta}})$. Turning to (2.), observe that (14) is a strictly convex function of $\widetilde{\boldsymbol{\beta}}$; this follows from the fact that $\sum_{j=1}^{p}(\tau_j|\beta_j| + \lambda\varepsilon\beta_j^2)$ is convex and that $-\ell(\widetilde{\boldsymbol{\beta}})$ is strictly convex and twice differentiable. In addition, the function $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) \geq 0$ is continuous in both $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\alpha}}$, twice continuously differentiable in $\widetilde{\boldsymbol{\beta}}$ for each $\widetilde{\boldsymbol{\alpha}}$, and has $h(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}) = 0$ when $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\alpha}}$ and is strictly positive otherwise. As a result, and in combination with (1.), (14) strictly majorizes $\xi(\widetilde{\boldsymbol{\beta}})$. Since all conditions of Theorem 2.1 are now satisfied, the stated convergence result for the MM algorithm immediately follows.

The result (3.) establishes the form of the update used at each iteration of the MM algorithm; its proof is an easy consequence of the results in Combettes and Wajs [13]. Observe that $\xi^{[\mathrm{S}]}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$ is separable in β_j for $j = 0, \ldots, p$; hence, minimization over $\mathbb{R}^{p+1}$ may be done component-wise. To be more precise, ignoring the leading term $-\ell(\tilde{\boldsymbol{\alpha}})$ in $\xi^{[\mathrm{S}]}(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}})$, the desired minimization problem corresponds to minimizing

$$f_0(\beta_0) + \sum_{j=1}^{p} \left(f_{1j}(\beta_j) + f_{2j}(\beta_j)\right),$$

where $f_0(\beta_0) = [-\nabla\ell(\tilde{\boldsymbol{\alpha}})]_0(\beta_0 - \alpha_0) + \varpi^{-1}(\beta_0 - \alpha_0)^2$ and, for $j \geq 1$, $f_{1j}(\beta_j) = \tau_j|\beta_j|$ and $f_{2j}(\beta_j) = [-\nabla\ell(\tilde{\boldsymbol{\alpha}})]_j\ (\beta_j - \alpha_j) + \varpi^{-1}(\beta_j - \alpha_j)^2 + \lambda\varepsilon\beta_j^2$.

Consider first β_j for $j \geq 1$. Observe that $f_{2j}(\beta_j)$ is twice continuously differentiable and strictly convex in β_j. Then, using Examples 2.16 and 2.20 and Proposition 3.1 of Combettes and Wajs [13], it follows immediately that the minimizer of $f_{1j}(\beta_j) + f_{2j}(\beta_j)$ is given by

$$\beta_j^* = \frac{1}{\zeta + 2\lambda\varepsilon} s\left([\nabla\ell(\tilde{\boldsymbol{\alpha}})]_j + \zeta\alpha_j, \tau_j\right) \tag{24}$$

where $\zeta = 2\varpi^{-1}$. Proceeding similarly, and noting that $f_0(\beta_0)$ is both strictly convex and twice continuously differentiable, we obtain the solution $\beta_0^* = ([\nabla\ell(\tilde{\boldsymbol{\alpha}})]_0 + \zeta\alpha_0)/\zeta$. This proves (15); a direct proof of these results is also provided in Schifano [57].

Turning to (16), take $\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\kappa}}$ for any bounded vectors $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^{(p+1)}$ and $\tilde{\boldsymbol{\kappa}} = (\kappa_0, \boldsymbol{\kappa}^T)^T \in \mathbb{R}^{(p+1)}$. Define $\gamma_j = \tilde{p}(|\alpha_j|; \boldsymbol{\lambda}_j) - \tilde{p}'(|\alpha_j|; \boldsymbol{\lambda}_j)|\alpha_j|$ for $j = 1, \ldots, p$. Then, following arguments similar to those in Daubechies, Defreise and De Mol

[16, Proposition 2.1], we may write

$$\xi^{[\mathrm{S}]}(\tilde{\boldsymbol{\beta}}+\tilde{\boldsymbol{\kappa}},\tilde{\boldsymbol{\alpha}}) = \xi^{[\mathrm{S}]}(\tilde{\boldsymbol{\beta}},\tilde{\boldsymbol{\alpha}}) + (\frac{\zeta}{2}+\lambda\varepsilon)\boldsymbol{\kappa}'\boldsymbol{\kappa} + \frac{\zeta}{2}\kappa_0^2 + \kappa_0(\zeta\beta_0 - \zeta\alpha_0 - [\nabla\ell(\tilde{\boldsymbol{\alpha}})]_0) + \sum_{j=1}^{p}\left[\tau_j(|\beta_j+\kappa_j| - |\beta_j|) + \kappa_j((d+2\lambda\varepsilon)\beta_j - \zeta\alpha_j - [\nabla\ell(\widetilde{\boldsymbol{\alpha}})]_j)\right].$$

Consider $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}^* \equiv [\beta_0^*, \boldsymbol{\beta}^{*T}]^T$ where $\tilde{\boldsymbol{\beta}}^*$ defined in (15), and define sets $\mathcal{J} = \{1,2,\ldots,p\}$, $\mathcal{J}_0 = \{j \in \mathcal{J} : \beta_j^* = 0\}$ and $\mathcal{J}_1 = \mathcal{J}\backslash\mathcal{J}_0$. Noting that β_j^* satisfies $(\zeta+2\lambda\varepsilon)\beta_j^* - \zeta\alpha_j - [\nabla\ell(\boldsymbol{\alpha})]_j = -\tau_j\mathrm{sign}(\beta_j^*)$ for $j \in \mathcal{J}_1$, and noting that $\zeta\beta_0^* - \zeta\alpha_0 - [\nabla\ell(\tilde{\boldsymbol{\alpha}})]_0 = 0$, we have (after some simplification)

$$\xi^{[\mathrm{S}]}(\tilde{\boldsymbol{\beta}}^*+\tilde{\boldsymbol{\kappa}},\tilde{\boldsymbol{\alpha}}) - \xi^{[\mathrm{S}]}(\tilde{\boldsymbol{\beta}}^*,\tilde{\boldsymbol{\alpha}}) = (\frac{\zeta}{2}+\lambda\varepsilon)\boldsymbol{\kappa}'\boldsymbol{\kappa} + \frac{\zeta}{2}\kappa_0^2 + \sum_{j\in\mathcal{J}_0}[\tau_j|\kappa_j| - \kappa_j(\zeta\alpha_j + [\nabla\ell(\boldsymbol{\alpha})]_j)] + \sum_{j\in\mathcal{J}_1}\left[\tau_j(|\beta_j^*+\kappa_j| - |\beta_j^*|) - \kappa_j\tau_j\mathrm{sign}(\beta_j^*)\right].$$

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

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

If $\beta_j^* < 0$, then

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

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

References

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

[2] Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York. MR1198884

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

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

[5] Antoniadis, A., Gijbels, I. and Nikolova, M. (2009). Penalized likelihood regression for generalized linear models with nonquadratic penalties. *Ann. Inst. Statist. Math.* 1-31. 10.1007/s10463-009-0242-4.

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

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

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

[9] Bohning, D. and Lindsay, B. (1988). Monotonocity of Quadratic-Approximation Algorithms. *Ann. Inst. Statist. Math.* **40** 641–663. MR0996690

[10] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press. MR2061575

[11] Chrétien, S. and Hero, A. O. (2008). On EM algorithms and their proximal generalizations. *ESAIM Journ. on Probability and Statistics* **12** 308–326. MR2404033

[12] Clarke, F. H. (1990). *Optimization and Nonsmooth Analysis.* SIAM, Philadelphia. MR1058436

[13] Combettes, P. L. and Wajs, V. R. (2005). Signal Recovery by Proximal Forward-Backward Splitting. *Multiscale Model. Simul.* **4** 1168-1200. MR2203849

[14] Cox, D. R. (1972). Regression models and life-tables (with Discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758 (49 ##6504)

[15] Cox, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509 (53 ##4340)

[16] Daubechies, I., Defreise, M. and De Mol, C. (2004). An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint. *Commun. Pure Appl. Math.* 1413–1457. MR2077704

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

[18] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

[19] Efron, B., Hastie, T., Johnstone, L. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–452. MR2060166

[20] Engler, D. and Li, Y. (2009). Survival analysis with high-dimensional covariates: an application in microarray studies. *Statist. Appl. Gen. Mol. Biol.* **8** Article 14.
[21] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.* **96** 1348-1360. MR1946581
[22] Fan, J., Feng, Y., Samworth, R. and Wu, Y. (2009a). SIS: Sure Independence Screening R package version 0.2.
[23] Fan, J., Samworth, R., and Wu, Y. (2009b). Ultrahigh dimensional variable selection: beyond the linear model. *Journal of Machine Learning Research* **10** 1829-1853. MR2550099
[24] Friedman, J., Hastie, T. and Tibshirani, R. (2008). Regularization Paths for Generalized Linear Models via Coordinate Descent Technical Report, Dept. of Statistics, Stanford University.
[25] Fygenson, M. and Ritov, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.* **22** 732–746. MR1292538
[26] Geman, D. and Reynolds, G. (1992). Constrained restoration and the recovery of discontinuities. *IEEE Trans. Patt. Anal. Mach. Intell* **14** 367–383.
[27] Gui, J. and Li, H. (2005a). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21** 3001–3008.
[28] Gui, J. and Li, H. (2005b). Threshold Gradient Descent Method for Censored Data Regression with Applications in Pharmacogenomics. *Pacific Symposium on Biocomputing* **10** 272–283.
[29] Hale, E. T., Yin, W. and Zhang, Y. (2008). Fixed-point Continuation for l1-minimization: Methodology and Convergence. *SIAM J. Optim.* **19** 1107–1130. MR2460734
[30] Hastie, T. and Efron, B. (2007). lars: Least Angle Regression, Lasso and Forward Stagewise R package version 0.9-7.
[31] Hiriart-Urruty, J. B. and Lemaréchal, C. (1996). *Convex Analysis and Minimization Algorithms I: Fundamentals.* Springer. MR1261420
[32] Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509
[33] Hunter, D. and Li, R. (2005). Variable Selection Using MM Algorithms. *Ann. Statist.* **33** 1617–1643. MR2166557
[34] Johnson, B. A., Lin, D.-Y. and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression problems. *J. Amer. Statist. Assoc.* **103** 672–680. MR2435469
[35] Johnson, L. M. and Strawderman, R. L. (2009). Induced smoothing for the semiparametric accelerated failure time model: asymptotics and extensions to clustered data. *Biometrika* **96** 577-590.
[36] Kim, Y., Choi, H. and Oh, H.-S. (2008). Smoothly Clipped Absolute Deviation on High Dimensions. *J. Amer. Statist. Assoc.* **103** 1665–1673. MR2510294

[37] Lange, K. (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. *J. Roy. Statist. Soc. Ser. B* **57** 425–437. MR1323348
[38] Lange, K. (2004). *Optimization.* Springer, New York, USA. MR2072899
[39] Lange, K., Hunter, D. and Yang, I. (2000). Optimization Transfer Using Surrogate Objective Functions. *J. of Comp. Graph. Statist.* **9** 1–20. MR1819865
[40] Li, H. and Gui, J. (2004). Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* **20** Suppl. 1, i208–215.
[41] Li, H. and Luan, Y. (2005). Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics* **21** 2403–2409.
[42] Luenberger, D. G. and Ye, Y. (2008). *Linear and nonlinear programming.*, Third ed. *International Series in Operations Research & Management Science, 116*. Springer, New York. MR2423726
[43] Ma, S. and Huang, J. (2007). Additive risk survival model with microarray data. *BMC Bioinformatics* **8** 192.
[44] Martinez-Climent, J. A., Alizadeh, A. A., Segraves, R., Blesa, D., Rubio-Moscardo, F., Albertson, D. G., Garcia-Conde, J., Dyer, M. J., Levy, R., Pinkel, D. and Lossos, I. S. (2003). Transformation of follicular lymphoma to diffuse large cell lymphoma is associated with a heterogeneous set of DNA copy number and gene expression alterations. *Blood* **101** 3109–3117.
[45] Mazumder, R., Friedman, J. and Hastie, T. (2009). SparseNet: Coordinate Descent with Non-Convex Penalties Technical Report, Department of Statistics, Stanford University.
[46] McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions.*, 2 ed. Wiley-Interscience. MR2392878
[47] Meyer, R. R. (1976). Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *J. Comput. System. Sci.* **12** 108–121. MR0439211
[48] Meyer, R. R. (1977). A comparison of the forcing function and point-to-set mapping approaches to convergence analysis. *SIAM J. Control Optimization* **15** 699–715. MR0472056
[49] Nikolova, M. (2000). Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.* **61** 633–658. MR1780806
[50] Ortega, and Rheinboldt, (1970). *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York. MR0273810
[51] Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69** 659–677. MR2370074
[52] Polak, E. (1987). On the mathematical foundations of nondifferentiable optimization in engineering design. *SIAM Rev.* **29** 21–89. MR880500 (88e:49063)
[53] Roland, C. and Varadhan, R. (2005). New iterative schemes for nonlinear fixed point problems, with applications to problems with bifurca-

tions and incomplete-data problems. *Appl. Numer. Math.* **55** 215–226. MR2160751

[54] Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., López-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T. and Staudt, L. M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346** 1937–1947.

[55] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35** 1012–1030. MR2341696

[56] Saeki, K., Miura, Y., Aki, D., Kurosaki, T. and Yoshimura, A. (2003). The B cell-specific major raft protein, Raftlin, is necessary for the integrity of lipid raft and BCR signal transduction. *EMBO J.* **22** 3015–3026.

[57] Schifano, E. D. (2010). Topics in Penalized Estimation PhD thesis, Cornell University.

[58] Sha, N., Tadesse, M. G. and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22** 2262–2268.

[59] She, Y. (2009). Thresholding-based Iterative Selection Procedures for Model Selection and Shrinkage. *Electronic Journal of Statistics* **3** 384-415. MR2501318

[60] Sinisi, S. E., Neugebauer, R. and van der Laan, M. J. (2006). Cross-validated bagged prediction of survival. *Statist. Appl. Gen. Mol. Biol.* **5** Article 12. MR2221294

[61] Sohn, I., Kim, J., Jung, S. H. and Park, C. (2009). Gradient lasso for Cox proportional hazards model. *Bioinformatics* **25** 1775–1781.

[62] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[63] Tibshirani, R. J. (2009). Univariate shrinkage in the cox model for high dimensional data. *Statist. Appl. Gen. Mol. Biol.* **8** Article21.

[64] Tseng, P. (2004). An analysis of the EM algorithm and entropy-like proximal point methods. *Mathematics of Operations Research* **29** 27–44. MR2065712

[65] Tseng, P. and Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming B* **117** 387–423. MR2421312

[66] Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statist. Sinica* **15** 831–840. MR2233916

[67] Varadhan, R. and Roland, C. (2008). Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm. *Scand. J. Statist.* **35** 335–353. MR2418745

[68] Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867

[69] Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.* **2** 224–244. MR2415601

[70] Zangwill, W. I. (1969). *Nonlinear Programming; a Unified Approach.* Prentice-Hall International Series in Management, Englewood Cliffs, N.J. MR0359816

[71] Zhang, C.-H. (2008). Discussion of "One-step sparse estimates in nonconcave penalized likelihood models" by H. Zou and R. Li. *Ann. Statist.* **36** 1553–1560. MR2435446

[72] Zhang, C.-H. (2010). Nearly Unbiased Variable Selection Under Minimax Concave Penalty. *Ann. Statist.* **38** 894–942. MR2604701

[73] Zhang, D. and Zhang, M. (2007). Bayesian profiling of molecular signatures to predict event times. *Theor. Biol. Med. Model.* **4** 3.

[74] Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

[75] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320. MR2137327

[76] Zou, H. and Hastie, T. (2008). elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA R package version 1.0-5.

[77] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

[78] Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **34** 1733-1751. MR2533470