

A Connection Between Variable Selection and EM-Type Algorithms

DAVID R. HUNTER AND RUNZE LI

Department of Statistics
Pennsylvania State University
University Park, PA 16802-2111
dhunter@stat.psu.edu
rli@stat.psu.edu

June 20, 2002

Abstract

Variable selection is fundamental to high-dimensional statistical modeling. Fan and Li (2001) proposed a class of variable selection procedures via nonconcave penalized likelihood. Optimizing the penalized likelihood function is challenging as it is a high-dimensional nonconcave function with singularities. A new algorithm is proposed for finding a solution of the nonconcave penalized likelihood via a modified local quadratic approximation. The proposed algorithm repairs the drawback of Fan and Li's algorithm. We establish a connection between local quadratic approximation and the so-called MM algorithms, useful extensions of the EM algorithms. This connection enables us to analyze the local and global convergence of the local quadratic approximation algorithm by employing the techniques used for EM algorithms. Moreover, this connection provides a general scheme for constructing a minorizing function in the MM algorithm via the local quadratic approximation.

Key Words: EM algorithm, LASSO, MM algorithm, penalized likelihood, oracle estimator, SCAD.

Abbreviated: Variable Selection and MM algorithms

1 Introduction

Fan and Li (2001) propose a new family of variable selection methods via a nonconcave penalized likelihood approach. This family includes many useful variable selection procedures, such as bridge regression (Frank and Friedman, 1993) and LASSO (Tibshirani, 1996). It has been shown that with a proper choice of the regularization parameter, the resulting estimates possess an oracle property, namely, they work as well as if the correct submodel were known. See Fan and Li (2001) for details. It is a very challenging task to optimize a nonconcave penalized likelihood, since the target function is a high-dimensional nonconcave function with singularities. Fan and Li (2001) propose a new and generic algorithm based on local quadratic approximation. In this paper, we demonstrate a connection between the local quadratic approximation algorithm and minorization-maximization (MM) algorithms (Lange et al., 2000). MM algorithms exploit an optimization technique that extends the central idea of EM algorithms (Dempster et al., 1977) to situations not necessarily involving missing data nor even maximum likelihood estimation. This connection enables us to analyze the convergence of the local quadratic approximation algorithm using the techniques related to EM algorithms (Wu, 1983; Meng, 1994; Lange 1995; Meng and Van Dyk, 1997).

The local quadratic approximation algorithm suffers from a drawback of forward variable selection: If a covariate is deleted at any step, it will be excluded from the final selected model. To repair this drawback, we propose a new algorithm based on a modification of local quadratic approximation. The newly proposed algorithm retains virtues of the Newton-Raphson algorithm and is numerically stable. This also enables us to compute the standard error for the resulting estimator via the sandwich formula. Further, the new algorithm is never forced to delete a covariate in the process of iteration. The general convergence results known for MM algorithms imply among other things that the newly proposed algorithm converges correctly to the maximizer of the penalized likelihood whenever this maximizer is the unique local maximum.

The rest of the paper is organized as follows. Section 2 briefly introduces the nonconcave penalized likelihood approach. After providing some background on MM algorithms, Section 3 investigates the convergence properties of local quadratic approximation using a

connection with MM algorithms, then gives a modification of the local quadratic approximation algorithm that remedies a drawback of the original algorithm. Section 4 describes some simulation studies, and possible extensions are discussed in Section 5. All technical proofs appear in the Appendix.

2 Nonconcave penalized likelihood and variable selection

Suppose that $\{(x_i, y_i), i = 1, \dots, n\}$ is a random sample with conditional log-likelihood $\ell_i(\beta) (\equiv \ell_i(x_i^T \beta, y_i))$ given x_i . As discussed in Fan and Li (2001), a form of penalized likelihood is

$$Q(\beta) = \sum_{i=1}^n \ell_i(\beta) - n \sum_{j=1}^d \lambda_j p_j(|\beta_j|) \equiv \ell(\beta) - n \sum_{j=1}^d \lambda_j p_j(|\beta_j|), \quad (2.1)$$

where the $p_j(\cdot)$ s are penalty functions, d is the dimension of the covariate vector x_i , and the λ_j s are tuning parameters controlling model complexity. Often, the λ_j s may be chosen by a data-driven approach such as cross-validation or generalized cross-validation (Craven and Wahba, 1979). The penalty function $p_j(\cdot)$ and the tuning parameters λ_j are not necessarily the same for all j . This allows one to incorporate hierarchical prior information for the unknown coefficients by using different penalty functions and taking different values of λ_j for the different regression coefficients. For instance, one may not be willing to penalize important factors in practice. For ease of presentation, we assume throughout this paper that the same penalization is applied to every component of β and write $\lambda_j p_j(|\beta_j|)$ as $p_\lambda(|\beta_j|)$, which implies that the penalty function is allowed to depend on λ . Extensions to situations with different penalty functions for each component of β do not involve any extra difficulties except more tedious notation.

Many variable selection criteria are special cases of the penalized likelihood of equation (2.1). For instance, consider the L_0 penalty $p_\lambda(\beta) = 0.5\lambda^2 I\{|\beta| \neq 0\}$, also called the entropy penalty in the literature, where $I(\cdot)$ is an indicator function. With this penalty, many variable selection criteria, such as AIC and BIC, can be written in the form of (2.1). Recently, many authors have been working on penalized least squares with the L_q penalty $p_\lambda(|\beta|) = \lambda|\beta|^q$. Indeed, bridge regression is the solution of penalized least squares with the L_q penalty (Frank and Friedman, 1993). It is well known that ridge regression is the

solution of penalized likelihood with the L_2 penalty. The L_1 penalty results in LASSO, proposed by Tibshirani (1996).

To achieve the purpose of variable selection, one might impose certain conditions on the penalty functions. Antoniadis and Fan (2001) and Fan and Li (2001) argue that a good penalty function should result in an estimator with the following three properties: **unbiasedness** for a large true coefficient to avoid excessive estimation bias, **sparsity** (estimating a small coefficient as zero) to reduce model complexity, and **continuity** to avoid unnecessary variation in model prediction. However, the aforementioned penalty functions do not simultaneously satisfy these three mathematical conditions, which require that $p_\lambda(\cdot)$ should be a concave function over $(0, \infty)$ and that $p'_\lambda(0+) > 0$. We refer to the latter condition as singularity at the origin.

The smoothly clipped absolute deviation (SCAD) penalty, used in Fan and Li (2001), whose first derivative is defined as

$$p'_\lambda(\beta) = \lambda I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{a - 1} I(\beta > \lambda) \quad \text{for some } a > 2 \text{ and } \beta > 0, \quad (2.2)$$

with $p_\lambda(0) = 0$, satisfies all three mathematical requirements.

2.1 Bayesian Interpretation

The penalized likelihood admits Bayesian interpretations. The penalty term in the penalized likelihood corresponds to a prior for β that is proportional to

$$\prod_{j=1}^d \exp\{-np_\lambda(|\beta_j|)\}.$$

For the L_2 penalty, the corresponding prior for β is a Gaussian distribution, while for the L_1 penalty, the prior corresponds to a double exponential distribution. For the L_0 and SCAD penalties, the corresponding prior distributions are improper. The prior for β in the current setup is somewhat unusual in that it depends on the sample size n . Figure 1 depicts some prior distributions. In Figure 1, all three prior distributions corresponding to the L_1 , $L_{0.5}$, and SCAD penalties are seen to be singular at the origin. This is a necessary condition for the resulting estimate to automatically reduce model complexity. Also notice that the shapes of the prior distributions corresponding to the L_1 , $L_{0.5}$ and SCAD penalties are convex over $[0, +\infty)$, which is different from the Gaussian distribution.

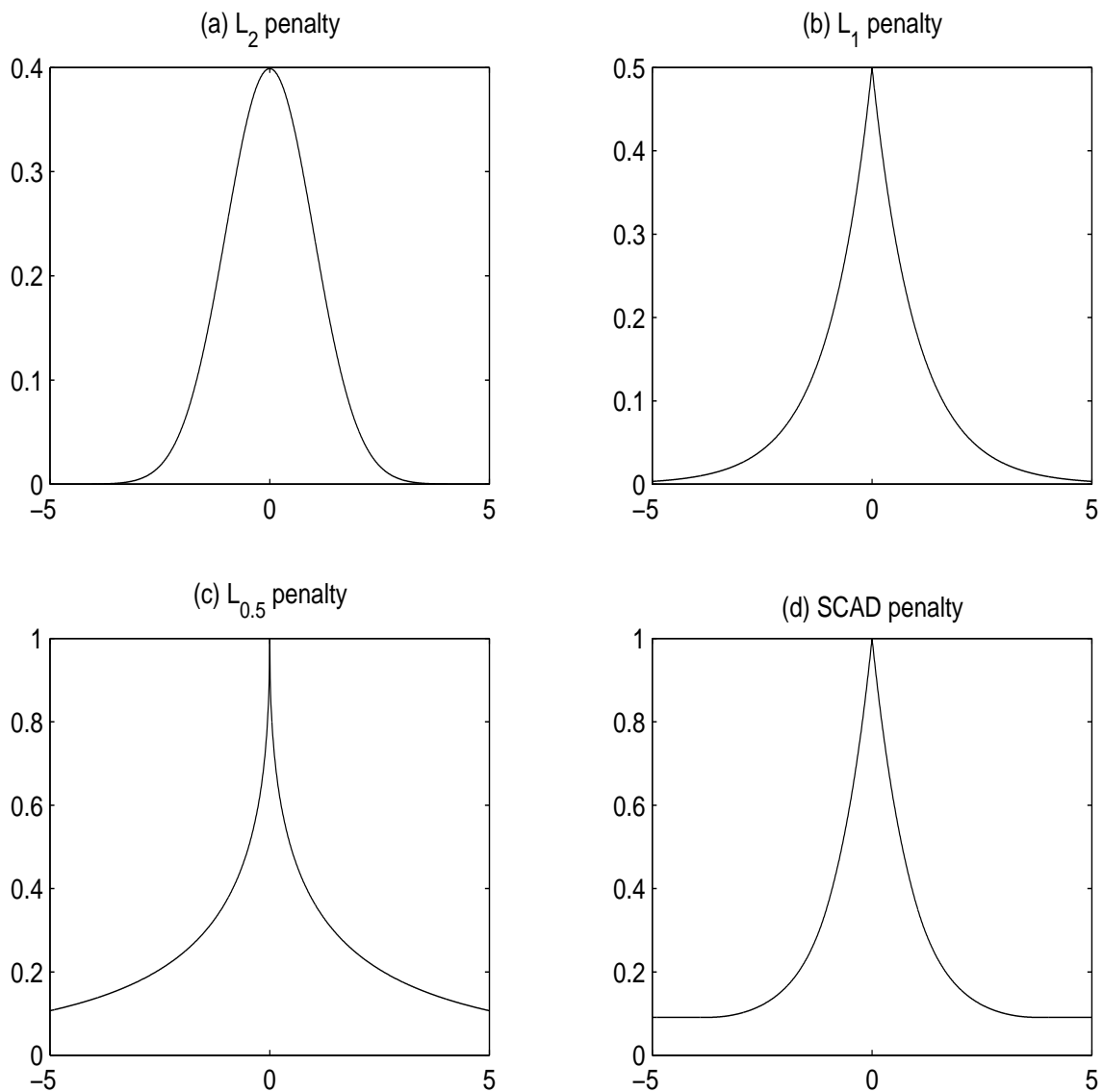


Figure 1: *Plots of prior distributions. (a)—(d) are plots of prior distributions corresponding to L_2 , L_1 , $L_{0.5}$ and SCAD penalties, respectively.*

One may introduce a more complicated prior distribution, but this is unnecessary. For example, as shown in Fan and Li (2001), with a proper choice of the tuning parameter and the SCAD penalty, the resulting estimator possesses an oracle property. Furthermore, the fact that the log-prior separates the parameter components β_j allows us to use a local quadratic approximation (see Section 3 below) for finding the solution of the penalized

likelihood, making computation issues easy to handle.

3 Connection between local quadratic approximation and MM algorithms

It is a challenging task to find the nonconcave penalized likelihood estimate. Fan and Li (2001) proposed the local quadratic approximation for the nonconcave penalty function: Suppose that we are given an initial value $\beta^{(0)}$ that is close to the true value of β . If $\beta_j^{(0)}$ is very close to 0, then set $\widehat{\beta}_j = 0$; otherwise, the penalty function is locally approximated by a quadratic function as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sgn}(\beta_j) \approx \{p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}\beta_j$$

when $\beta_j \neq 0$. In other words,

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{1}{2}\{p'_\lambda(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2}) \quad (3.1)$$

for $\beta_j \approx \beta_{j0}$. With the aid of the local quadratic approximation, a Newton-Raphson algorithm (for example) can be used to maximize the penalized likelihood function, where each iteration updates the local quadratic approximation.

In this section, we show that this local quadratic approximation idea is an instance of an MM algorithm. This fact enables us to study the convergence properties of the algorithm using techniques applicable to MM algorithms in general.

3.1 Local Quadratic Approximation as an MM algorithm

MM stands for Majorize-Minimize or Minorize-Maximize, depending on the context (Lange et al., 2000). EM algorithms (Dempster et al., 1977) are the most famous examples of MM algorithms, though there are many examples of MM algorithms that involve neither maximum likelihood nor missing data. Heiser (1995) and Lange et al. (2000) give partial surveys of work in this area. The apparent ambiguity in allowing MM to stand for two different things is harmless, since any maximization problem may be viewed as a minimization problem by changing the sign of the objective function.

Consider the penalty term

$$-n \sum_{j=1}^d p_\lambda(|\beta_j|)$$

of equation (2.1), ignoring its minus sign for the moment. Mimicking the idea of Equation (3.1), we define the function

$$\Phi_{\theta_0}(\theta) = p_\lambda(|\theta_0|) + \frac{(\theta^2 - \theta_0^2)p'_\lambda(|\theta_0|)}{2|\theta_0|}. \quad (3.2)$$

Note that $\Phi_{\theta_0}(\cdot)$ is undefined when $\theta_0 = 0$; this issue will be addressed in section 3.2. We are interested in penalty functions $p_\lambda(\theta)$ for which

$$\Phi_{\theta_0}(\theta) \geq p_\lambda(\theta) \text{ for all } \theta, \text{ with equality when } \theta = \theta_0. \quad (3.3)$$

A function $\Phi_{\theta_0}(\theta)$ satisfying condition (3.3) is said to *majorize* $p_\lambda(\theta)$ at θ_0 (if the direction of the inequality in condition (3.3) were reversed, then $\Phi_{\theta_0}(\theta)$ would be said to *minorize* $p_\lambda(\theta)$ at θ_0).

The driving force behind an MM algorithm is the fact that condition (3.3) implies

$$\Phi_{\theta_0}(\theta) - \Phi_{\theta_0}(\theta_0) \geq p_\lambda(\theta) - p_\lambda(\theta_0),$$

which in turn gives the *descent property*

$$\Phi_{\theta_0}(\theta) < \Phi_{\theta_0}(\theta_0) \text{ implies } p_\lambda(\theta) < p_\lambda(\theta_0). \quad (3.4)$$

In other words, if θ_0 denotes the current iterate, any decrease in the value of $\Phi_{\theta_0}(\theta)$ guarantees a decrease in the value of $p_\lambda(\theta)$. A minimization algorithm would exploit this fact by repeatedly constructing the majorizing function $\Phi_{\theta_k}(\theta)$, then minimizing it to give θ_{k+1} —hence the name “majorize-minimize algorithm”—where k is the iteration number.

It remains to verify that condition (3.3) holds for the particular choice of $\Phi_{\theta_0}(\theta)$ given in equation (3.2) for $\theta_0 \neq 0$. This is easily seen to be the case whenever $p'_\lambda(\theta)/\theta$ is a nonincreasing function on $(0, \infty)$ since $\Phi_{\theta_0}(\theta_0) = p_\lambda(\theta_0)$ and

$$\frac{d}{d\theta} [\Phi_{\theta_0}(\theta) - p_\lambda(\theta)] = \theta \left[\frac{p'_\lambda(|\theta_0|)}{|\theta_0|} - \frac{p'_\lambda(|\theta|)}{|\theta|} \right], \quad (3.5)$$

which implies that the even function $\Phi_{\theta_0}(\theta) - p_\lambda(\theta)$ is decreasing on $(0, |\theta_0|)$ and increasing on $(|\theta_0|, \infty)$. Condition (3.3) is satisfied for many penalty functions in the literature, including

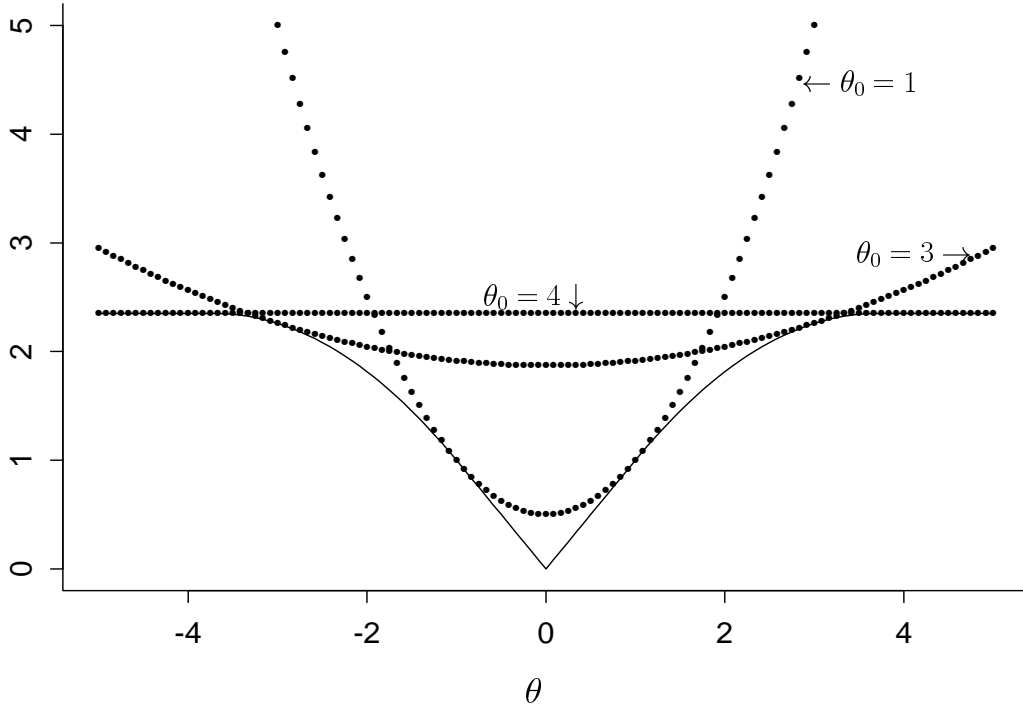


Figure 2: Majorizing functions for the SCAD penalty function. The solid curve is $p_\lambda(\theta)$ with $a = 3.7$ and $\lambda = 1$; the dotted curves are $\Phi_{\theta_0}(\theta)$ for three different values of θ_0 .

all penalty functions considered in this paper. The SCAD penalty function and its majorizer $\Phi_{\theta_0}(\theta)$ are depicted in Figure 2.

Let $\beta^{(k)} = (\beta_1^{(k)}, \dots, \beta_d^{(k)})$ denote the value of β at the k th iteration of the iterative algorithm. Suppose that none of the $\beta_j^{(k)}$ equals zero, so that $\Phi_{\beta_j^{(k)}}(\beta_j)$ is defined for all j . Then because we have demonstrated that $\Phi_{\beta_j^{(k)}}(\beta_j)$ majorizes $p_\lambda(\beta_j)$ at $\beta_j^{(k)} \neq 0$, clearly

$$S_k(\beta) \equiv \ell(\beta) - n \sum_{j=1}^d \Phi_{\beta_j^{(k)}}(\beta_j) \quad (3.6)$$

minorizes $Q(\beta)$ at $\beta^{(k)}$. By the *ascent property*—the obvious analogue, for minorizing functions, of the descent property (3.4)—we know that we should define $\beta^{(k+1)}$ so that $S_k(\beta^{(k+1)}) > S_k(\beta^{(k)})$ if possible, thus ensuring that $Q(\beta^{(k+1)}) > Q(\beta^{(k)})$. The benefit of essentially replacing one maximization problem by another in this way is that $S_k(\beta)$ is sus-

ceptible to a gradient-based scheme such as Newton-Raphson, unlike the nondifferentiable $Q(\beta)$. Note in particular that the form of the penalty term in equation (3.6) is extremely simple—it is the sum of quadratic functions of the scalars β_j —meaning that the difficulty in maximizing $S_k(\beta)$ is determined solely by the form of $\ell(\beta)$. For example, in the special case of a linear regression model with normally distributed errors, the loglikelihood function $\ell(\beta)$ is quadratic, which implies that $S_k(\beta)$ may be maximized analytically.

If some of the components of $\beta^{(k)}$ equal zero (or in practice, if some of them are close to zero), the algorithm proceeds by simply setting the final estimates of those components to be zero, deleting them from consideration, then defining the function $S_k(\beta_N)$ as in equation (3.6), where β_N is the vector composed of the nonzero components of β . The drawback of this scheme is that once a component is set to zero, it may never reenter the model at a later stage of the algorithm. In Section 3.2, we present a version of the algorithm that does not suffer from this drawback.

3.2 An improved version of local quadratic approximation

The problem with the MM algorithm defined in the preceding section is that the majorizing function $\Phi_{\theta_0}(\theta)$ of equation (3.2) is not defined for $\theta_0 = 0$. The discontinuity at the origin of the derivatives of certain penalty functions, such as SCAD or L_1 , makes it impossible to construct a quadratic majorizing function at $\theta_0 = 0$.

Intuitively, it seems that the problem of dividing by zero could be solved if $\Phi_{\theta_0}(\theta)$ were replaced by

$$\Phi_{\theta_0, \epsilon}(\theta) = p_\lambda(|\theta_0|) + \frac{(\theta^2 - \theta_0^2)p'_\lambda(|\theta_0|)}{2(\epsilon + |\theta_0|)} \quad (3.7)$$

for some small $\epsilon > 0$, where we take $p'_\lambda(|\theta_0|) = p'_\lambda(|\theta_0|+)$ if $\theta_0 = 0$. However, this perturbed version of $\Phi_{\theta_0}(\theta)$ is no longer a majorizer of $p_\lambda(\theta)$ as required by the MM theory. Thus, we pursue a strategy of defining a slightly perturbed objective function $Q_\epsilon(\beta)$, then maximizing it using an MM algorithm. The constant ϵ may be eventually lowered to zero once the iterates near convergence.

For a given penalty function $p_\lambda(\theta)$ such that the following integral exists and is finite, we define

$$p_{\lambda, \epsilon}(\theta) = p_\lambda(|\theta|) - \epsilon \int_0^{|\theta|} \frac{p'_\lambda(t)}{\epsilon + t} dt \quad (3.8)$$

for small $\epsilon > 0$, and

$$Q_\epsilon(\beta) = \ell(\beta) - n \sum_{j=1}^d p_{\lambda,\epsilon}(|\beta_j|). \quad (3.9)$$

Theorem 3.1 *Suppose that $p'_\lambda(\theta)$ is nonincreasing on $(0, \infty)$. Then*

(a) *If $p'_\lambda(0+) < \infty$, then $|Q_\epsilon(\beta) - Q(\beta)| \rightarrow 0$ uniformly on compact subsets of the parameter space as $\epsilon \rightarrow 0$.*

(b) *For any given $\epsilon > 0$,*

$$S_{k,\epsilon}(\beta) \equiv \ell(\beta) - n \sum_{j=1}^d \Phi_{\beta_j^{(k)},\epsilon}(\beta_j) \quad (3.10)$$

minorizes $Q_\epsilon(\beta)$ at $\beta^{(k)}$.

The proof of Theorem 3.1 is given in the Appendix. Although $S_{k,\epsilon}(\beta)$ does not minorize $Q(\beta)$, from Theorem 3.1, it minorizes $Q_\epsilon(\beta)$ and $Q_\epsilon(\beta)$ is close to $Q(\beta)$ provided that ϵ is small enough. Thus, intuitively, the maximizer of $Q_\epsilon(\beta)$ should be close to the maximizer of $Q(\beta)$ as long as ϵ is small and $Q(\beta)$ is not too flat in the neighborhood of the maximizer. Indeed, when the algorithm converges, it follows by straightforward differentiation of $S_{k,\epsilon}(\beta)$ that

$$\frac{\partial S_{k,\epsilon}(\hat{\beta})}{\partial \beta_j} = \frac{\partial \ell(\hat{\beta})}{\partial \beta_j} - n p'_\lambda(|\hat{\beta}_j|) \operatorname{sgn}(\hat{\beta}_j) \frac{1}{\epsilon |\hat{\beta}_j|^{-1} + 1} = 0.$$

Thus, as $\epsilon \downarrow 0$, the resulting estimator satisfies the penalized likelihood equation:

$$\frac{\partial \ell(\hat{\beta})}{\partial \beta_j} - n \operatorname{sgn}(\hat{\beta}_j) p'_\lambda(|\hat{\beta}_j|) = 0.$$

For a given $\epsilon > 0$, we maximize $Q_\epsilon(\beta)$ using an MM algorithm without having to ignore components for which $\beta_j = 0$. Then ϵ may be gradually decreased to zero. Any β_j that remain zero after this procedure are presumably true zeros, since each component has been given the chance to “escape” zero while $\epsilon > 0$. In the simulation studies of Section 4, we take ϵ to be a small multiple of the standard error of the ordinary maximum likelihood estimate.

3.3 The algorithm

By the ascent property of an MM algorithm, it is desirable to construct $\beta^{(k+1)}$ at the k th iteration so that

$$S_{k,\epsilon}(\beta^{(k+1)}) > S_{k,\epsilon}(\beta^{(k)}). \quad (3.11)$$

To get more insight into how to construct a sequence of $\beta^{(k)}$ such that inequality (3.11) holds, we start with linear regression models with normal random errors. For ease of presentation, assume that the variance σ^2 of the random error is known. Thus, after dropping a constant term,

$$\ell(\beta) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - x_i^T \beta)^2$$

and therefore, $S_{k,\epsilon}(\beta)$ is a quadratic function of β , which means that $S_{k,\epsilon}(\beta)$ is maximized exactly by

$$\beta^{(k+1)} = \{X^T X + nE_k\}^{-1} X^T y, \quad (3.12)$$

where $X = (x_1, \dots, x_n)^T$, the design matrix of the linear regression model, y consists of y_i , and $E_k = E(\beta^{(k)})$ with $E(\beta) = \text{diag}\{e_1(\beta_1), \dots, e_d(\beta_d)\}$ and $e_j(\beta_j) = p'_\lambda(|\beta_j|)/(\epsilon + |\beta_j|)$. The algorithm defined by equation (3.12) may be viewed as iterative ridge regression.

For generalized linear models with the canonical link, the Hessian matrix of the log-likelihood function is

$$\ddot{\ell}(\beta) = -\sum_{i=1}^n v(x_i^T \beta) x_i x_i^T,$$

where $v(\cdot)$ is the variance function, which implies that the Hessian matrix is negative definite provided that X is of full rank. Furthermore, under some mild regularity conditions, it can be shown that for any β satisfying $\|\beta - \beta_0\| = o_P(1)$,

$$\ell(\beta) = \ell(\beta_0) + (\beta - \beta_0)^T \dot{\ell}(\beta_0) + \frac{1}{2}(\beta - \beta_0)^T \ddot{\ell}(\beta_0)(\beta - \beta_0) \{1 + o_P(1)\}, \quad (3.13)$$

where β_0 is the true value of β and $\dot{\ell}(\beta)$ stands for $\partial \ell(\beta) / \partial \beta$.

We take the ordinary maximum likelihood estimate to be the initial value $\beta^{(0)}$ in our algorithm. Again under some regularity conditions, the ordinary maximum likelihood estimate exists. Furthermore, the ordinary maximum likelihood estimate is root- n consistent for the true value of β_0 ; i.e., $\|\beta^{(0)} - \beta_0\| = O_P(n^{-1/2})$. Therefore (3.13) holds. Thus,

with probability tending to one, the Newton-Raphson algorithm yields a sequence of $\beta^{(k)}$ satisfying the inequality (3.11).

In more general cases, the Hessian matrix $\ddot{\ell}(\beta)$ may not be negative definite or it may be difficult to compute. It is possible to substitute the Fisher information matrix instead. Under the regularity conditions for local asymptotic normality (see, for example, Lehmann, 1983), it follows that for any β satisfying $\|\beta - \beta_0\| = o_P(1)$,

$$\ell(\beta) = \ell(\beta_0) + (\beta - \beta_0)^T \dot{\ell}(\beta_0) - \frac{1}{2}(\beta - \beta_0)^T I(\beta_0)(\beta - \beta_0) \{1 + o_P(1)\}, \quad (3.14)$$

where $I(\beta_0)$ is the Fisher information matrix, which is finite positive definite under certain conditions. Starting once again from the ordinary maximum likelihood estimate and using a modified Newton-Raphson algorithm that uses $-I(\beta)$ rather than the Hessian matrix, we can construct a sequence of $\beta^{(k)}$ satisfying inequality (3.11) with probability tending to one.

The Newton-Raphson algorithm enables a standard error estimate via a sandwich formula:

$$\text{cov}(\hat{\beta}) \approx \{\ddot{\ell}(\hat{\beta}) - nE_k\}^{-1} \text{cov}\{\dot{S}_{k,\epsilon}(\hat{\beta})\} \{\ddot{\ell}(\hat{\beta}) - nE_k\}^{-1}. \quad (3.15)$$

Naturally, another estimate may be formed if $-I(\hat{\beta})$ is substituted for $\ddot{\ell}(\hat{\beta})$.

The above discussion is based on asymptotic properties of the log-likelihood function. In practice, particularly when the sample size n is small, we may modify the Newton-Raphson algorithm using a step-halving scheme. After k iterations of the algorithm, the search direction is $\Delta_k = \{\ddot{\ell}(\beta^{(k)}) - nE_k\}^{-1} \dot{S}_{k,\epsilon}(\beta^{(k)})$. This search direction is multiplied by some $\alpha_k \in (0, 1]$ chosen so that

$$S_{k,\epsilon}(\beta^{(k)} + \alpha_k \Delta_k) > S_{k,\epsilon}(\beta^{(k)}). \quad (3.16)$$

To find α_k , we first set $\alpha_k = 1$ and then check inequality (3.16). If the inequality does not hold, we replace α_k by $\alpha_k/2$ and check again, repeating the process until inequality (3.16) holds. If $\ddot{\ell}(\beta^{(k)}) - nE_k$ is negative definite, this strategy is guaranteed to find α_k to satisfy (3.16). If the information matrix is used in place of the Hessian, then $-I(\hat{\beta}) - nE_k$ is guaranteed to be negative definite. When (3.16) is satisfied, we set $\beta^{(k+1)} = \beta^{(k)} + \alpha_k \Delta_k$ and then repeat the whole process until some convergence criterion is reached.

Note that the step-halving scheme above does not actually maximize $S_{k,\epsilon}(\beta)$. It would be a waste of time to do so; once a suitable α_k is found to satisfy inequality (3.16), it is

better simply to set $\beta^{(k+1)} = \beta^{(k)} + \alpha_k \Delta_k$ and move on to construct $S_{k+1,\epsilon}(\beta)$ at the next iteration. The reason for this is that the rate of convergence of the algorithm is governed primarily by the quality of the approximation of $Q_\epsilon(\beta)$ by $S_{k,\epsilon}(\beta)$ and is not very dependent on whether $S_k(\beta)$ is maximized at each iteration. See Lange (1995) for details.

3.4 Convergence

It is not possible to prove that a generic MM algorithm converges at all, and when an MM algorithm does converge, there is no guarantee that it converges to the global maximum. For example, there are well-known pathological examples in which EM algorithms converge to saddle points or fail to converge (McLachlan and Krishnan, 1997). Nonetheless, it is often possible to obtain convergence results in specific cases.

We define a stationary point of the function $Q_\epsilon(\beta)$ to be any point β at which the gradient vector is zero. Because the differentiable function $S_{k,\epsilon}(\beta)$ is tangent to $Q_\epsilon(\beta)$ at the point $\beta^{(k)}$ by the minorization property, the gradient vectors of $S_{k,\epsilon}(\beta)$ and $Q_\epsilon(\beta)$ are equal when evaluated at $\beta^{(k)}$. Thus, when using the method of Section 3.3 to maximize $S_{k,\epsilon}(\beta)$, we see that fixed points of the algorithm— i.e., points with gradient zero— coincide with stationary points of $Q_\epsilon(\beta)$. Letting $M(\beta)$ denote the map implicitly defined by the algorithm that takes $\beta^{(k)}$ to $\beta^{(k+1)}$ for any point $\beta^{(k)}$, we showed that $S_{k,\epsilon}\{M(\beta)\} > S_{k,\epsilon}(\beta)$ whenever $\dot{S}_{k,\epsilon}(\beta) \neq \mathbf{0}$. By the ascent property of the MM algorithm, we conclude $Q_\epsilon\{M(\beta)\} \geq Q_\epsilon(\beta)$, with equality only if β is a stationary point. Similar to Lyapunov’s theorem in Lange (1995), we have the following proposition.

Proposition 3.1 *Given an initial value $\beta^{(0)}$, let $\beta^{(k)} = M^k(\beta^{(0)})$. If $Q_\epsilon(\beta) = Q_\epsilon\{M(\beta)\}$ only for stationary points β of Q_ϵ , then any limit point β^* of the sequence $\{\beta^{(k)}\}$ is also a stationary point of $Q_\epsilon(\beta)$ if $M(\beta)$ is continuous at β^* .*

A brief proof of Proposition 3.1 is given in the Appendix. From Proposition 3.1, we see that if $Q_\epsilon(\beta)$ has at most one stationary point—for example, if it is strictly concave—then any limit point must be the unique stationary point. Furthermore, in such a case we may be assured that the algorithm will converge to the unique global maximizer as long as it is possible to prove that the iterates of the algorithm are confined to some compact set, since

that implies the existence of a limit point.

4 Numerical Comparisons

Our simulation studies use generalized cross-validation (GCV, Craven and Wahba, 1979) to select the tuning parameters λ . As suggested by Fan and Li (2001), we take $a = 3.7$ in the definition of SCAD.

Example 1. In this example, we generated 500 data sets, each of which consists of 40 observations from the model

$$y = x^T \beta + \sigma \varepsilon,$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, and the components of x and ε are standard normal. The correlation between x_i and x_j is $0.5^{|i-j|}$. This is a model used in Tibshirani (1996) and Fan and Li (2001). These authors have compared the performance of the SCAD and the LASSO with other existing ones. Here we focus on comparisons of the proposed algorithms for the SCAD. We also present comparisons of the performance of the SCAD and the best subset variable selection procedures. We used the values 9, 3, and 1 for σ and compared the performance of the proposed procedures in terms of model error (ME), defined as $\text{ME}(\hat{\beta}) = (\hat{\beta} - \beta)^T \text{cov}(x)(\hat{\beta} - \beta)$, and model complexity. The ME's of the proposed procedures are divided by that of the ordinary maximum likelihood (least squares) estimate. The mean and standard deviation of the relative model error (RME) over 500 simulated data sets are summarized in Table 1. The average number of 0 coefficients is also reported in Table 1, in which the column labeled ‘‘C’’ gives the average number of coefficients, of the five true zeros, correctly set to zero and the column labeled ‘‘I’’ gives the average number of the three true nonzeros incorrectly set to zero.

In Tables 1 and 2, the rows labeled ‘‘SCAD’’ give results of penalized likelihood with the SCAD penalty, using Fan and Li’s (2001) local quadratic approximation algorithm. The ‘‘ModSCAD1’’ rows present simulation results of the penalized likelihood with the SCAD penalty using the modified local quadratic algorithm proposed in Section 3. In our simulations, we used a different ϵ for each β_j , taken to be 10^{-5} times the standard error of the ordinary likelihood estimate for that parameter; these ϵ were never lowered to

Table 1: Relative model errors for 500 simulated data sets

Method	RME			Zeros			RME			Zeros		
	mean (std)	C	I	mean (std)	C	I	mean (std)	C	I	mean (std)	C	I
	$\sigma = 9$			$\sigma = 3$			$\sigma = 1$					
SCAD	0.731(0.503)	4.29	1.56	0.753(0.510)	4.39	0.25	0.463(0.237)	4.45	0			
ModSCAD1	0.687(0.459)	2.78	0.82	0.742(0.472)	3.37	0.11	0.468(0.235)	3.49	0			
ModSCAD2	0.691(0.467)	3.80	1.25	0.743(0.475)	3.93	0.17	0.467(0.235)	3.99	0			
BS(BIC)	0.870(0.565)	4.55	1.84	0.752(0.565)	4.59	0.33	0.518(0.279)	4.62	0			
BS(AIC)	0.871(0.409)	4.01	1.47	0.768(0.347)	4.01	0.19	0.687(0.254)	4.04	0			
Oracle	0.358(0.230)	5.00	0	0.358(0.230)	5.00	0	0.358(0.230)	5.00	0			

zero. The resulting estimate was set to zero if it was less than 10^{-5} . The “ModSCAD2” rows differ from the “ModSCAD1” rows only in their criterion for setting a parameter estimate to zero; in these rows, coefficients were set to zero when less than 0.1 standard error of the ordinary maximum likelihood estimate in absolute value. BS(AIC) and BS(BIC) stand for the best subset variable selection procedures that minimize AIC scores and BIC scores. Finally, “Oracle” stands for the oracle estimate computed by using the true model $y = \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \varepsilon$.

The case with $\sigma = 9$ corresponds to a high noise level, with standard errors of roughly 2.1 for all the ordinary least squares estimates. This case is very challenging to all procedures because the sizes of $\beta_1 = 3$, $\beta_2 = 1.5$, and $\beta_5 = 2$ are not large compared to their standard errors. From the left panel of Table 1, it can be seen that the modified local quadratic approximation improves Fan and Li’s local quadratic approximation by reducing model error and reducing the number of coefficients erroneously set to zero. This is consistent with our expectation in this case where several of the true coefficients are small but not zero because the modified local quadratic algorithm was proposed to remedy a drawback of the original local quadratic approximation: If a covariate is deleted at any step, it will be excluded from the final selected model. There seems to be a price to pay for this benefit in terms of model complexity, as evidenced by the “C” column, but the modified versions did not reduce the values of ϵ to zero as they could have.

The case with $\sigma = 3$ corresponds to a moderate noise level, with standard errors of

roughly 0.7 for all the ordinary least squares estimates. This is challenging to all procedures, since $\beta_2 = 1.5$ is only about twice its standard error. From Table 1, the performance of all variable selection procedures are similar in terms of model error.

The case $\sigma = 1$ corresponds to a low noise level for which all variable selection procedures are expected to work well. From Table 1, SCAD and its modifications outperform the best subset variable selection procedures. The performance of the modified local quadratic algorithm is similar to Fan and Li's local quadratic algorithm. This is because nonzero coefficients were not incorrectly set to be zero using the original version of the local quadratic algorithm, which is evidenced from the last column in Table 1. From the right panel of Table 1, ModSCAD2 slightly improves upon ModSCAD1 in terms of reduction of model complexity. This is also true for the other two cases. From Table 1, the performance of SCAD nears that of the oracle estimator as the noise level decreases.

Table 2: Standard deviations and standard errors of $\hat{\beta}_1$ when $\sigma = 1$

Method	SD	SE (std(SE))	90% Coverage	95% Coverage
SCAD	0.2021	0.1912(0.0427)	0.876	0.948
ModSCAD1	0.2021	0.1905(0.0425)	0.876	0.934
ModSCAD2	0.2021	0.1905(0.0425)	0.876	0.934
BS(BIC)	0.2036	0.1858(0.0296)	0.918	0.964
BS(AIC)	0.2074	0.1851(0.0296)	0.912	0.964
Oracle	0.2003	0.1890(0.0297)	0.884	0.938

We now test the accuracy of the proposed standard error formula. The standard deviation, denoted by SD, of the estimated coefficients for the 500 simulated data sets can be regarded as the true standard error except for Monte Carlo error. The average of the estimated standard errors, denoted by SE, for the 500 simulated data sets, and the standard deviation, denoted by std(SE), of the estimated standard errors gauge the overall performance of the standard error formula. Table 2 only presents the SD, SE, std(SE) and the 90% and 95% coverage probabilities of β_1 when $\sigma = 1$. The results for other coefficients and values of σ are similar. From Table 2, the differences between SD and SE are less than one standard deviation of the SE, which suggests that the proposed standard error

formula works fairly well. However, the SE appears to consistently underestimate the SD, a common phenomenon (see Kauermann and Carroll, 2001), so it may benefit from some slight modification.

5 Extensions

We have demonstrated a connection between variable selection and MM algorithms and shown how to modify the MM algorithms so that they do not mistakenly eliminate variables too soon. Although we have discussed a very broad range of potential applications, using penalized likelihood with various likelihood functions and various penalty functions, the potential applicability of these ideas is even greater. For example, one can certainly imagine other penalty functions that are not discussed here. More importantly, these ideas can be extended beyond the realm of penalized likelihood, since there is very little in this paper that is particular to the likelihood setting. For example, penalized least squares and penalized partial likelihood (Fan and Li, 2002) may benefit from the algorithms described in this article.

Appendix: Proofs of results in Section 3

Proof of Theorem 3.1

For part (a), it is sufficient to show that $|p_{\lambda,\epsilon}(\theta) - p_{\lambda}(\theta)| \rightarrow 0$ uniformly on compact subsets of the parameter space as $\epsilon \rightarrow 0$. Since $p'_{\lambda}(\theta)$ is nonincreasing on $(0, \infty)$ and $p'_{\lambda}(0+) < \infty$,

$$|p_{\lambda,\epsilon}(\theta) - p_{\lambda}(\theta)| \leq \epsilon \log \left[1 + \frac{|\theta|}{\epsilon} \right] p'_{\lambda}(0+),$$

and clearly the right side of the above inequality tends to 0 uniformly on compact subsets of the parameter space as $\epsilon \rightarrow 0$.

To prove part (b), we show that $\Phi_{\theta_0,\epsilon}(\theta)$ defined in (3.7) majorizes $p_{\lambda,\epsilon}(\theta)$ at θ_0 . It follows by definition that $\Phi_{\theta_0,\epsilon}(\theta_0) = p_{\lambda,\epsilon}(\theta_0)$. Furthermore,

$$\frac{d}{d\theta} [\Phi_{\theta_0,\epsilon}(\theta) - p_{\lambda,\epsilon}(\theta)] = \theta \left[\frac{p'_{\lambda}(|\theta_0|)}{\epsilon + |\theta_0|} - \frac{p'_{\lambda}(|\theta|)}{\epsilon + |\theta|} \right].$$

Note that $p'_{\lambda}(\theta)$ is nonincreasing on $(0, \infty)$. Hence, $p'_{\lambda}(\theta)/(\epsilon + |\theta|)$ is nondecreasing, which

implies that the even function $\Phi_{\theta_0, \epsilon}(\theta) - p_{\lambda, \epsilon}(\theta)$ is decreasing on $(0, |\theta_0|)$ and increasing on $(|\theta_0|, \infty)$, giving the desired result.

Proof of Proposition 3.1.

Given an initial value $\beta^{(0)}$, let $\beta^{(k)} = M^k(\beta^{(0)})$ for $k \geq 1$; i.e., $\{\beta^{(k)}\}$ is the sequence of points that the MM algorithm generates starting from $\beta^{(0)}$. Let Λ denote the set of limit points of this sequence. For any $\beta^* \in \Lambda$, passing to a subsequence we have $\beta^{(k_n)} \rightarrow \beta^*$. The quantity $Q_\epsilon(\beta^{(k_n)})$, since it is increasing in n and bounded above, converges to a limit as $n \rightarrow \infty$. Thus, taking limits in the inequalities

$$Q_\epsilon(\beta^{(k_n)}) \leq Q_\epsilon\{M(\beta^{(k_n)})\} \leq Q_\epsilon(\beta^{(k_{n+1})})$$

gives $Q_\epsilon(\beta^*) = Q_\epsilon\{\lim_{n \rightarrow \infty} M(\beta_{k_n})\}$, assuming this limit exists. Of course, if $M(\beta)$ is continuous at β^* , then we have $Q_\epsilon(\beta^*) = Q_\epsilon\{M(\beta^*)\}$, which implies that β^* is a stationary point of $Q_\epsilon(\beta)$.

Note that Λ is not necessarily nonempty in the above proof. However, we know that each $\beta^{(k)}$ lies in the set $\{\beta : Q_\epsilon(\beta) \geq Q_\epsilon(\beta_1)\}$, so if this set is compact, as is often the case, we may conclude that Λ is indeed nonempty.

References

- Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussion). *Journal of American Statistical Association*, **96**, 939-967.
- Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377-403.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, **96**, 1348-1360.

- Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, **30**, 74-99.
- Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.
- Heiser, W.J. (1995). Convergent computing by iterative majorization: theory and applications in multidimensional data analysis. In *Recent Advances in Descriptive Multivariate Analysis* (ed. W.J. Krzanowski), pp. 157–189. Clarendon Press, Oxford.
- Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **57**, 425–437.
- Lange, K. Hunter, D.R. and Yang, I. (2000). Optimization transfer using surrogate objective functions, *Journal of Computational and Graphical Statistics*, **9**, 1–59.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. Pacific Grove, California: Wadsworth & Brooks/Cole.
- Meng, X.-L. (1994). On the rate of convergence of the ECM algorithm. *The Annals of Statistics*, **22**, 326-339
- Meng, X.-L. and Van Dyk, D. A. (1997). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 511–567.
- McLachlan, G. and Krishnan, T. (1997). *The EM Algorithm and Extensions*. Wiley, New York.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, series B*, **58**, 267-288.
- Wu, C. F. J. (1983). On the convergence of properties of the EM algorithm, *The Annals of Statistics*, **11**, 95-103.