Statistica Sinica **26** (2016), 903-923
doi:http://dx.doi.org/10.5705/ss.202014.0011

THE Mnet METHOD FOR VARIABLE SELECTION

Jian Huang[1], Patrick Breheny[1], Sangin Lee[2], Shuangge Ma[3] and Cun-Hui Zhang[4]

[1] University of Iowa, [2] UT Southwestern Medical Center, [3] Yale University and [4] Rutgers University

Abstract: We propose a penalized approach for variable selection using a combination of minimax concave and ridge penalties. The method is designed to deal with $p \geq n$ problems with highly correlated predictors. We call this the Mnet method. Similar to the elastic net of Zou and Hastie (2005), the Mnet tends to select or drop highly correlated predictors together. However, unlike the elastic net, the Mnet is selection consistent and equal to the oracle ridge estimator with high probability under reasonable conditions. We develop an efficient coordinate descent algorithm to compute the Mnet estimates. Simulation studies show that the Mnet has better performance in the presence of highly correlated predictors than either the elastic net or MCP. We illustrate the application of the Mnet to data from a gene expression study in ophthalmology.

Key words and phrases: Correlated predictors, minimax concave penalty, oracle property, $p > n$ problems, ridge regression.

1. Introduction

There has been much work on penalized methods for variable selection and estimation in high-dimensional regression models. Several important methods have been proposed, including estimators based on the bridge penalty (Frank and Friedman (1993)), the ℓ_1 penalty or least absolute shrinkage and selection operator (LASSO, Tibshirani (1996)), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)), and the minimax concave penalty (MCP, Zhang (2010)). These methods provide a computationally feasible way to carry out variable selection in high-dimensional settings and much progress has been made in understanding the theoretical properties.

While these methods have many attractive properties, they have some drawbacks. For example, as pointed out by Zou and Hastie (2005), for a linear regression model with p predictors and sample size n, the LASSO can select at most n variables; it tends to only select one variable among a group of highly correlated variables; and its prediction performance is not as good as the ridge regression if there exists high correlation among predictors. To overcome these limitations, Zou and Hastie (2005) proposed the elastic net (Enet), which uses a combination of the ℓ_1 and ℓ_2 penalties. Yuan and Lin (2007) obtained a result for the Enet to select the true model in the classical settings when p is fixed. Jia and Yu (2010) studied the selection consistency property of the Enet estimator when $p \gg n$. They showed that under an irrepresentable condition and certain other conditions, the Enet is selection consistent. Their results generalize those of Zhao and Yu (2006) on the selection consistency of the LASSO under the irrepresentable condition. But the Enet estimator is asymptotically biased because of the ℓ_1 component in the penalty and it cannot achieve selection consistency and estimation efficiency simultaneously. Zou and Zhang (2009) proposed the adaptive Enet estimator and provided sufficient conditions under which it is oracle. However, they require that the singular values of the design matrix be uniformly bounded away from zero and infinity. Thus their results excludes the case of highly correlated predictors and are only applicable to the situations when $p < n$.

There is a need to develop methods that are applicable to $p \geq n$ regression problems with highly correlated predictors and have the oracle property. Inspired by the Enet and MCP methodologies, we propose a penalized approach that uses a combination of the MCP and ℓ_2 penalty. We call this the Mnet. Similar to the Enet, the Mnet can effectively deal with highly correlated predictors in $p \geq n$ situations. It encourages a grouping effect in selection, meaning that it selects or drops highly correlated predictors together. Because the Mnet uses the MCP instead of the ℓ_1 penalty for selection, it has important advantages. The Mnet is selection consistent under a sparse Riesz condition on the 'ridge design matrix', which only requires a submatrix of this matrix to be nonsingular. This condition is usually less restrictive than the irrepresentable condition, especially in high-dimensional settings (Zhang (2010)). The Mnet estimator is equal to the oracle ridge estimator with high probability, in the sense that it correctly selects predictors with nonzero coefficients and estimates the selected coefficients using ridge regression.

This article is organized as follows. In Section 2, we define the Mnet estimator and discuss its basic characteristics. In Section 3, we present a coordinate descent algorithm for computing the estimates. Results on the sign consistency of Mnet and its equivalency to the oracle ridge estimator are presented in Section 4. In Section 5, we conduct simulation studies to evaluate its finite sample performance and illustrate its application using a real data example. Final remarks are given in Section 6. Proofs are provided in the Supplementary Materials.

2. The Mnet Estimator

Consider the linear regression model

$$y = \sum_{j=1}^{p} x_j \beta_j + \varepsilon, \tag{2.1}$$

where $y = (y_1, \ldots, y_n)'$ is the vector of n response variables, $x_j = (x_{1j}, \ldots, x_{nj})'$ is the jth predictor vector, β_j is the jth regression coefficient and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the vector of random errors. Let $X = (x_1, \ldots, x_p)$ be the design matrix. We assume that the responses are centered and the covariates are centered and standardized, so that the intercept term is zero and $n^{-1} \sum_{i=1}^{n} x_{ij}^2 = 1$.

2.1. Definition

We first provide a brief description of the MCP introduced by Zhang (2010). The MCP is given by

$$\rho(t; \lambda_1, \gamma) = \lambda_1 \int_0^{|t|} (1 - \frac{x}{\gamma\lambda_1})_+ dx, \tag{2.2}$$

where λ_1 is a penalty parameter and γ is a regularization parameter that controls the concavity of ρ. We require $\lambda_1 \geq 0$ and $\gamma > 1$. Here $x_+ = x1_{\{x \geq 0\}}$. The MCP can be easily understood by considering its derivative, which is

$$\dot{\rho}(t; \lambda_1, \gamma) = \lambda_1 \big(1 - \frac{|t|}{\gamma\lambda_1}\big)_+ \mathrm{sgn}(t), \tag{2.3}$$

where $\mathrm{sgn}(t) = -1, 0$, or 1 if $t < 0, = 0$, or > 0. It begins by applying the same rate of penalization as the lasso, but continuously relaxes that penalization until, when $|t| > \gamma\lambda_1$, the rate of penalization drops to 0. It provides a continuum of penalties with the ℓ_1 penalty at $\gamma = \infty$ and the hard-thresholding penalty as $\gamma \to 1+$.

For $\lambda = (\lambda_1, \lambda_2)$ with $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$, consider the penalized criterion

$$M(b; \lambda, \gamma) = \frac{1}{2n} \|y - Xb\|^2 + \sum_{j=1}^{p} \rho(|b_j|; \lambda_1, \gamma) + \frac{1}{2}\lambda_2 \sum_{j=1}^{p} b_j^2, \quad b \in \mathbb{R}^p. \tag{2.4}$$

The Enet criterion uses the ℓ_1 penalty in the first penalty term while here we use the MCP. For a given (λ, γ), the Mnet estimator is defined as

$$\hat{\beta}_{Mnet}(\lambda, \gamma) = \underset{b}{\mathrm{argmin}}\, M(b; \lambda, \gamma). \tag{2.5}$$

Our rationale for using the MCP in (2.4) is as follows. As discussed in Fan and Li (2001), a good penalty function should result in an estimator with three basic properties: unbiasedness, sparsity and continuity. The ℓ_1 penalty produces estimators that are sparse and continuous with respect to data, but are biased. To remove the bias in the estimators resulting from the ℓ_1 penalty and to achieve oracle efficiency, they proposed the SCAD penalty for variable selection and estimation. In an in-depth analysis of the LASSO, SCAD, and MCP, Zhang (2010) showed that they belong to the family of quadratic spline penalties with sparsity and continuity properties. The MCP is the simplest penalty that results in an estimator that is nearly unbiased, sparse and continuous. Further discussions on the advantages of the MCP over other popular penalties can be found in Mazumder, Friedman, and Hastie (2011).

2.2. Orthonormal designs

To gain some insights into the characteristics of the Mnet estimator, consider the case where the design matrix is orthonormal. In this case, the problem simplifies to estimation in p univariate models of the form

$$y_i = x_{ij}\theta + \varepsilon_i,\ 1 \le i \le n.$$

Let $z = n^{-1}\sum_{i=1}^{n} x_{ij}y_i$ be the least squares estimator of θ (since $n^{-1}\sum_{i=1}^{n} x_{ij}^2 = 1$). The corresponding Mnet criterion can be written as

$$\frac{1}{2}(z-\theta)^2 + \rho(\theta;\lambda_1,\gamma) + \frac{1}{2}\lambda_2\theta^2. \tag{2.6}$$

When $\gamma(1+\lambda_2) > 1$, the minimizer $\hat{\theta}_{Mnet}$ of (2.6) is

$$\hat{\theta}_{Mnet} = \begin{cases} \text{sgn}(z)\frac{\gamma(|z|-\lambda_1)_+}{\gamma(1+\lambda_2)-1} & \text{if } |z| \le \gamma\lambda_1(1+\lambda_2), \\ \frac{z}{1+\lambda_2} & \text{if } |z| > \gamma\lambda_1(1+\lambda_2). \end{cases} \tag{2.7}$$

This expression illustrates a key feature of the Mnet estimator. In most of the sample space of z, it is the same as the ridge estimator. Specifically, for small $\gamma\lambda_1(1+\lambda_2)$, the probability of the region where $\hat{\theta}_{Mnet}$ is not equal to the ridge estimator is also small. In Section 4, we show that this remains true for general designs under reasonable conditions.

It is instructive to compare the Mnet with the Enet. The naive Enet (nEnet) estimator is

$$\hat{\theta}_{nEnet} = \underset{\theta}{\text{argmin}}\ \frac{1}{2}(z-\theta)^2 + \lambda_1|\theta| + \frac{1}{2}\lambda_2\theta^2 = \text{sgn}(z)\frac{(|z|-\lambda_1)_+}{1+\lambda_2}.$$

The ridge penalty introduces an extra bias factor $1/(1+\lambda_2)$. This ridge shrinkage on top of the LASSO shrinkage is the double shrinkage effect discussed in Zou and Hastie (2005). They proposed to remove the ridge shrinkage factor by multiplying the naive Enet by $(1+\lambda_2)$ to obtain the Enet estimator

$$\tilde{\theta}_{Enet} = (1+\lambda_2)\hat{\theta}_{nEnet} = \text{sgn}(z)(|z|-\lambda_1)_+.$$

Thus for orthonormal designs, the (rescaled) Enet estimator is the same as the LASSO estimator and is still biased.

Similarly, we can rescale $\hat{\theta}_{Mnet}$ to obtain the re-scaled Mnet estimator, written as

$$\tilde{\theta}_{sMnet} = \begin{cases} \frac{\gamma(1+\lambda_2)}{\gamma(1+\lambda_2)-1}\hat{\theta}_{Enet} & \text{if } |z| \le \gamma\lambda_1(1+\lambda_2), \\ z & \text{if } |z| > \gamma\lambda_1(1+\lambda_2). \end{cases}$$

It is equal to the unbiased estimator z when $|z| > \gamma\lambda_1(1+\lambda_2)$. As $\gamma(1+\lambda_2) \to \infty$, the Mnet converges to the Enet; as $\gamma(1+\lambda_2) \to 1$, the Mnet converges to the hard thresholding rule.

For orthogonal designs, re-scaling removes the bias due to the ridge shrinkage without significantly inflating the variance. However, it can be demonstrated numerically that for correlated designs, rescaling can substantially inflate the variance of the Mnet estimator and as a result, the mean squared error is increased. Since we focus on the variable selection property of the Mnet and rescaling does not affect selection results, we will not consider rescaling here.

2.3. Grouping effect

Similar to the Enet, the Mnet has a grouping effect. It tends to select or drop strongly correlated predictors together, due to the ℓ_2 penalty term. For simplicity, we write $\hat{\beta}_j$ for $\hat{\beta}_{Mnet,j}$.

Proposition 1. *Let $\rho_{jk} = n^{-1}\sum_{i=1}^n x_{ij}x_{ik}$ be the correlation coefficient between x_j and x_k. Suppose $\lambda_2 > 0$. If*

$$\xi = \begin{cases} \max\{2\gamma(\gamma\lambda_2 - 1)^{-1}, (\gamma\lambda_2 + 1)(\lambda_2(\gamma\lambda_2 - 1))^{-1}, \lambda_2^{-1}\} & \textit{if } \gamma\lambda_2 > 1, \\ \lambda_2^{-1} & \textit{if } \gamma\lambda_2 \le 1, \end{cases} \tag{2.8}$$

for $\rho_{jk} \ge 0$, we have

$$|\hat{\beta}_j - \hat{\beta}_k| \le \xi n^{-1/2}\sqrt{2(1-\rho_{jk})}\|y\|;$$

for $\rho_{jk} < 0$, we have

$$|\hat{\beta}_j + \hat{\beta}_k| \le \xi n^{-1/2}\sqrt{2(1+\rho_{jk})}\|y\|.$$

From this proposition, we see that the difference between $\hat{\beta}_j$ and $\hat{\beta}_k$ is bounded by a quantity determined by the correlation coefficient. It shows that highly correlated predictors tend to be selected together by the Mnet. In particular, $\hat{\beta}_j - \hat{\beta}_k \to 0$ as $\rho_{jk} \to 1$ and $\hat{\beta}_j + \hat{\beta}_k \to 0$ as $\rho_{jk} \to -1$.

3. Computation

3.1. The coordinate descent algorithm

We use the cyclical coordinate descent algorithm originally proposed for such criteria with convex penalties as the LASSO (Fu (1998), Friedman et al. (2007), Wu and Lange (2008)). It has been proposed to calculate the MCP estimates Breheny and Huang (2011). The algorithm optimizes a target function with respect to a single parameter at a time, iteratively cycling through all parameters until convergence is reached. It is particularly suitable for problems that have a simple closed form solution in a single dimension but lack one in higher dimensions. The authors have implemented the algorithm described here in the `R` package `ncvreg`, publicly available at `http://cran.r-project.org/web/packages/ncvreg`.

The problem, then, is to minimize M with respect to β_j, given current values for the regression coefficients $\tilde{\beta}_k$. Take

$$M_j(\beta_j;\lambda,\gamma)=\frac{1}{2n}\sum_{i=1}^{n}(y_i-\sum_{k\neq j}x_{ik}\tilde{\beta}_k-x_{ij}\beta_j)^2+\rho(|\beta_j|;\lambda_1,\gamma)+\frac{1}{2}\lambda_2\beta_j^2.$$

Let $\tilde{y}_{ij}=\sum_{k\neq j}x_{ik}\tilde{\beta}_k, \tilde{r}_{ij}=y_i-\tilde{y}_{ij},$ and $\tilde{z}_j=n^{-1}\sum_{i=1}^{n}x_{ij}\tilde{r}_{ij}$, where $\tilde{r}_{ij}$s are the partial residuals with respect to the j^{th} covariate. Some algebra shows that

$$M_j(\beta_j;\lambda,\gamma)=\frac{1}{2}(\beta_j-\tilde{z}_j)^2+\rho(|\beta_j|;\lambda_1,\gamma)+\frac{1}{2}\lambda_2\beta_j^2+\frac{1}{2n}\sum_{i=1}^{n}\tilde{r}_{ij}^2-\frac{1}{2}\tilde{z}_j^2.$$

Thus, if $\tilde{\beta}_j$ denotes the minimizer of $M_j(\beta_j;\lambda,\gamma)$, (2.6) and (2.7) imply that

$$\tilde{\beta}_j=\begin{cases}\operatorname{sgn}(\tilde{z}_j)\frac{\gamma(|\tilde{z}_j|-\lambda_1)_+}{\gamma(1+\lambda_2)-1} & \text{if } |\tilde{z}_j|\leq\gamma\lambda_1(1+\lambda_2)\\ \frac{\tilde{z}_j}{1+\lambda_2} & \text{if } |\tilde{z}_j|>\gamma\lambda_1(1+\lambda_2)\end{cases}, \tag{3.1}$$

for $\gamma(1+\lambda_2)>1$.

Given the current value $\tilde{\beta}^{(s)}$ in the sth iteration for $s=0,1\ldots$, the algorithm for determining $\hat{\beta}$ is as follows.

(1) Calculate

$$\tilde{z}_j=n^{-1}\sum_{i=1}^{n}x_{ij}\tilde{r}_{ij}=n^{-1}\sum_{i=1}^{n}x_{ij}(y_i-\tilde{y}_i+x_{ij}\tilde{\beta}_j^{(s)})=n^{-1}\sum_{i=1}^{n}x_{ij}r_i+\tilde{\beta}_j^{(s)},$$

where $\tilde{y}_i=\sum_{j=1}^{p}x_{ij}\tilde{\beta}_j^{(s)}$ is the current fitted value for observation i and $r_i=y_i-\tilde{y}_i$ is the current residual. The calculation of $\tilde{z}_j$ is carried out using the last expression in this equation.

(2) Update $\tilde{\beta}_j^{(s+1)}$ using (3.1).

(3) Update $r_i\leftarrow r_i-(\tilde{\beta}_j^{(s+1)}-\tilde{\beta}_j^{(s)})x_{ij}$ for all i.

The last step ensures that r_i's always hold the current values of the residuals. These three steps loop over all values of j and proceed iteratively until convergence. The coordinate descent algorithm has the potential to be extremely efficient, in that the above three steps require only $O(n)$ operations, meaning that one full iteration can be completed at a computational cost of $O(np)$ operations.

3.2. Pathwise optimization

Usually, we are interested in determining $\hat{\beta}$ for a range of values of (λ, γ), thereby producing a path of coefficient values through the parameter space. Consider the reparametrization: $\tau = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1/\tau$. Using it, we can compute solutions for decreasing values of τ, starting at the smallest value $\tau_{\max}$ for which all coefficients are 0 and continuing down to a minimum value $\tau_{\min}$, thereby obtaining the unique coefficient path for which the ratio between λ_1 and λ_2 is held constant at $\alpha/(1-\alpha)$. If $p < n$ and the design matrix is full rank, $\tau_{\min}$ can be 0. In other settings, the model can become excessively large or cease to be identifiable for small τ; in such cases, a value such as $\tau_{\min} = 0.01\tau_{\max}$ is appropriate.

From (2.7), $\tau_{\max} = \max_{1 \le j \le p} |n^{-1}x_j'y|/\alpha$. Starting at this value, for which $\hat{\beta}$ has the closed form solution 0, and proceeding along a continuous path ensures that the initial values are reasonably close to the solution for all points along the path, thereby improving the stability and efficiency of the algorithm.

3.3. Convexity of the objective function

The preceding remarks concerning unique solutions and continuous coefficient paths are only guaranteed for convex objective functions. Because the MCP is nonconvex, this is not always the case for the Mnet objective function; it is possible, however, for the convexity of the ridge penalty and the least-squares loss function to overcome the nonconvexity of the MCP and produce a convex objective function.

Proposition 2. *Let* $c_{\min}$ *denote the minimum eigenvalue of* $n^{-1}X'X$. *Then the objective function defined by* (2.4) *is a convex function of* b *on* $\mathbb{R}^p$ *if and only if* $\gamma > 1/(c_{\min} + \lambda_2)$.

This establishes the condition necessary for global convexity on $\mathbb{R}^p$. In $p \gg n$ settings, where highly sparse solutions are desired, we may be concerned only with convexity in the local region of the parameter space consisting of the covariates estimated to have nonzero coefficients. In this case, the above condition can be relaxed by considering the minimum eigenvalue of $n^{-1}X_A'X_A$ instead, where X_A is a modified design matrix consisting of only those columns for which $\hat{\beta}_j \neq 0$. The issue of local convexity is explored in greater detail in Breheny and Huang (2011).

4. Selection Properties

In this section, we study the selection properties of the Mnet estimator $\hat{\beta}_{Mnet}$ at (2.5). We provide sufficient conditions under which the Mnet estimator is sign consistent and equals the oracle ridge estimator.

For simplicity of notation, we write $\hat{\beta} = \hat{\beta}_{Mnet}$. Let $\Sigma = n^{-1}X'X$ and, for any $A \subseteq \{1, \ldots, p\}$, take $X_A = (x_j, j \in A)$, $\quad \Sigma_A = \frac{1}{n}X_A'X_A$. Let the true value of the regression coefficient be $\beta^o = (\beta_1^o, \ldots, \beta_p^o)'$. Denote by $O = \{j : \beta_j^o \neq 0\}$, the oracle set of indices of the predictors with nonzero coefficients in the underlying model. Let $\beta_*^o = \min\{|\beta_j|, j \in \mathcal{O}\}$ and set $\beta_*^o = \infty$ if $\mathcal{O}$ is empty. Denote the cardinality of $\mathcal{O}$ by $|\mathcal{O}|$ and let $d^o = |\mathcal{O}|$. Define the oracle ridge estimator by

$$\hat{\beta}^o(\lambda_2) = \underset{b}{\operatorname{argmin}}\{\frac{1}{2n}\|y - Xb\|^2 + \frac{1}{2}\lambda_2\|b\|^2, b_j = 0, j \notin \mathcal{O}\}. \tag{4.1}$$

It is not a feasible estimator, as the oracle set is unknown.

4.1. The $p < n$ case

We first consider the selection property of the Mnet estimator for the $p < n$ case. We require the following basic condition.

(A1) (a) The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with $\mathrm{E}\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$; (b) For any $x > 0$, $\mathrm{P}(|\varepsilon_i| > x) \leq K\exp(-Cx^\alpha), i = 1, \ldots, n$, where C and K are positive constants and $1 \leq \alpha \leq 2$.

Let $c_{\min}$ be the smallest eigenvalue of Σ, and let c_1 and c_2 be the smallest and largest eigenvalues of $\Sigma_{\mathcal{O}}$, respectively.

Set

$$\lambda_n = \alpha_n \frac{\sigma \log^{1/\alpha}(p - d^o + 1)}{\sqrt{n}} \quad \text{and} \quad \tau_n = \alpha_n \frac{\sigma\sqrt{c_2}\log^{1/\alpha}(d^o + 1)}{\sqrt{n}(c_1 + \lambda_2)}, \tag{4.2}$$

where $\alpha_n = 1$ if $1 < \alpha \leq 2$ and $\alpha_n = \log n$ if $\alpha = 1$. For error terms with double exponential tails, there is an extra $\log n$ factor in these expressions.

Theorem 1. *Suppose* (A1) *holds and* $\gamma > 1/(c_{\min} + \lambda_2)$. *If*

$$\beta_*^o > \gamma\lambda_1 + \frac{2\lambda_2\|\beta^o\|}{(c_1 + \lambda_2)} \text{ and } \lambda_1 > \frac{2\lambda_2\sqrt{c_2}\|\beta^o\|}{(c_1 + \lambda_2)}, \tag{4.3}$$

then $\mathrm{P}\big(sgn(\hat{\beta}) \neq sgn(\beta^o) \text{ or } \hat{\beta} \neq \hat{\beta}^o\big) \leq \pi_1 + \pi_2$, *where the sgn function applies to a vector componentwise and*

$$\pi_1 = \frac{2K_1\lambda_n}{\lambda_1} \quad \text{and } \pi_2 = \frac{2K_1\tau_n}{(\beta_*^o - \gamma\lambda_1)}. \tag{4.4}$$

Here K_1 *is a positive constant that depends only on the tail behavior of the error distribution in* (A1b).

We note that the upper bound on the probability of selection error is nonasymptotic. The condition $\gamma > 1/(c_{\min} + \lambda_2)$ ensures that the Mnet criterion is strictly convex so that the resulting estimate is unique. This condition also essentially restricts $c_{\min} > 0$, which can only be satisfied when $p < n$. The first inequality in (4.3) requires the nonzero coefficients not to be too small in order for the Mnet estimator to be able to distinguish nonzero from zero coefficients. The second inequality in (4.3) requires that λ_1 should be at least in the same order as λ_2. This condition indicates that there is a trade-off between the grouping effect and good theoretical properties. If λ_2 is too big, the Mnet estimator is not selection consistent due to the bias introduced by the ridge penalty; if λ_2 is too small, the grouping effect is diminished.

Corollary 1. *Suppose that the conditions of Theorem* 1 *are satisfied. If* $\lambda_1 \geq a_n\lambda_n$ *and* $\beta_*^o \geq \gamma\lambda_1 + a_n\tau_n$ *for* $a_n \to \infty$ *as* $n \to \infty$*, then* $\mathrm{P}(sgn(\hat{\beta}) \neq sgn(\beta^o)$ *or* $\hat{\beta} \neq \hat{\beta}^o) \to 0$.

By Corollary 1, $\hat{\beta}$ behaves like the oracle ridge estimator and has the same sign as the underlying regression coefficients with probability tending to one.

4.2. The $p \geq n$ case

We now consider the selection property of the Mnet estimator when $p \geq n$. In this case, the model is not identifiable without further conditions, since the design matrix X is always singular. However, if the model is sparse and the design matrix satisfies the sparse Riesz condition, or SRC (Zhang and Huang (2008)), then the model is identifiable and selection consistency can be achieved.

Let

$$\tilde{X} = \begin{pmatrix} X \\ \sqrt{n\lambda_2}\, I_p \end{pmatrix},$$

where I_p is a $p \times p$ identity matrix. This can be considered an 'enlarged design matrix' from ridge regularization. The jth column of $\tilde{X}$ is $\tilde{x}_j = (x_j', \sqrt{n\lambda_2}e_j')'$, where e_j is the jth unit vector in $\mathbb{R}^p$. For $A \subseteq \{1, \ldots, p\}$, take

$$\tilde{X}_A = (\tilde{x}_j, j \in A), \tilde{P}_A = \tilde{X}_A(\tilde{X}_A'\tilde{X}_A)^{-1}\tilde{X}_A'. \tag{4.5}$$

Denote the cardinality of A by $|A|$. We say that $\tilde{X}$ satisfies the sparse Reisz condition (SRC) with rank d^* and spectrum bounds $\{c_* + \lambda_2, c^* + \lambda_2\}$ if

$$0 < c_* + \lambda_2 \leq \frac{1}{n}\|\tilde{X}_A u\|_2^2 \leq c^* + \lambda_2 < \infty,\ \forall A \text{ with } |A| \leq d^*, u \in \mathbb{R}^{|A|}, \|u\| = 1, \tag{4.6}$$

where c_* and c^* satisfy

$$0 \leq c_* \leq \frac{1}{n}\|X_A u\|_2^2 \leq c^*,\ \forall A \text{ with } |A| \leq d^*, u \in \mathbb{R}^{|A|}, \|u\| = 1.$$

We allow either $c_* = 0$ or $\lambda_2 = 0$, but require $c_* + \lambda_2 > 0$. Below, we simply say that $\tilde{X}$ satisfies the SRC$(d^*, c_* + \lambda_2, c^* + \lambda_2)$ if (4.6) holds.

(A2) The matrix $\tilde{X}$ satisfies the SRC$(d^*, c_* + \lambda_2, c^* + \lambda_2)$, where d^* satisfies $d^* \geq d^o(K_* + 1)$ with $K_* = (c^* + \lambda_2)/(c_* + \lambda_2) - 1/2$, and d^o is the number of nonzero coefficients.

Let $m = d^* - d^o$. Write

$$\lambda_n^* = \alpha_n \frac{\sigma \log^{1/\alpha}(p - d^o + 1)}{\sqrt{n}} \sqrt{c^*} m_\alpha \max\Big\{1, \frac{\sqrt{c^*}}{m\sqrt{n}(c_* + \lambda_2)^2}\Big\}, \tag{4.7}$$

where $m_\alpha = 1$ if $\alpha = 2$ and $= m^{1/\alpha}$ if $1 \leq \alpha < 2$. Let π_1 and π_2 be as in (4.4). Set

$$\pi_1^* = K_1 \lambda_n^* / \lambda_1 \quad \text{and} \quad \pi_3 = K_1 \alpha_n \frac{8\sigma c^* \lambda_2 \sqrt{d^o} \log^{1/\alpha}(d^o + 1)}{mn(c_* + \lambda_2)}. \tag{4.8}$$

Theorem 2. *Suppose that* (A1) *and* (A2) *hold. If*

$$\gamma \geq (c_* + \lambda_2)^{-1} \sqrt{4 + \frac{c_* + \lambda_2}{c^* + \lambda_2}}, \tag{4.9}$$

$\lambda_1 > 2\lambda_2\sqrt{c_2}\|\beta^o\|/(c_1 + \lambda_2)$ *and* $\beta_*^o > \gamma\lambda_1 + 2\lambda_2\|\beta^o\|/(c_1 + \lambda_2)$,

$$\mathrm{P}\big(sgn(\hat{\beta}) \neq sgn(\beta^o) \ or \ \hat{\beta} \neq \hat{\beta}^o\big) \leq \pi_1 + \pi_1^* + \pi_2 + \pi_3.$$

Corollary 2. *Suppose that the conditions of Theorem* 2 *are satisfied. If* $\lambda_1 \geq a_n \lambda_n^*$ *and* $\beta_*^o \geq \gamma\lambda_1 + a_n \tau_n$ *for* $a_n \to \infty$ *as* $n \to \infty$, *then*

$$\mathrm{P}\big(sgn(\hat{\beta}) \neq sgn(\beta^o) \ or \ \hat{\beta} \neq \hat{\beta}^o\big) \to 0 \quad as \ n \to \infty.$$

Theorem 2 and Corollary 2 provide sufficient conditions for sign consistency and the oracle property of the Mnet estimator in $p \geq n$ situations. Again, the probability bound on the selection error in Theorem 2 is nonasymptotic. Comparing with Theorem 1, here the extra terms π_1^* and π_3 in the probability bound come from the need to reduce the original p-dimensional problem to a d^*-dimensional problem. Condition (4.9) ensures that the Mnet criterion is locally convex in any d^*-dimensional subspace. It is stronger than the minimal sufficient condition $\gamma > 1/(c_* + \lambda_2)$ for local convexity. This reflects the difficulty and extra efforts needed in reducing the dimension from p to d^*. The SRC in (A2) guarantees that the model is identifiable in any lower d^*-dimensional space, which contains the d^o-dimensional space of the underlying model, since $d^* > d^o$. The difference $d^* - d^o = K_* d^o$ depends on K_*, which is determined by the spectrum bounds in the SRC. In the proof of Theorem 2 given in the Supplementary Materials, the first crucial step is to show that the dimension of the Mnet estimator is bounded by d^* with high probability. Then the original p-dimensional problem reduces to a d^*-dimensional problem. The other conditions of Theorem 2 imply that the conditions of Theorem 1 are satisfied for $p = d^*$. After dimension reduction is achieved, we can use the same argument as in Theorem 1 to show sign consistency. The role of λ_n^* is similar to λ_n in (4.2). However, the expression of λ_n^* has an extra term that which arises from the need to reduce the dimension from p to d^*. If $1 < \alpha \leq 2$, c_* is bounded away from zero and c^* is bounded by a finite constant, then for sufficiently large n we have $\lambda_n^* = \lambda_n \sqrt{c^*}$. Finally, We note that our results allow $c_* \to 0$ and $c^* \to \infty$ as long as the conditions in Theorem 2 are satisfied. Thus Theorem 2 and Corollary 2 are applicable to models with highly correlated predictors. Finally, we allow $p \gg n$ in Theorem 2 Corollary 2. For example, in the simplest case of an error distribution with sub-gaussian tails $(\alpha = 2)$ and $\sqrt{c^*}/(m\sqrt{n}(c_* + \lambda_2)^2) \leq 1$ in (4.7) for sufficiently large n, we can have $p - d^o = \exp(o(n))$, where $o(n)/n \to 0$.

5. Simulation Studies

In principle, the Mnet estimator has three parameters one may consider tuning: τ, α, and γ, with $\tau = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1/\tau$. However, optimizing the performance of the method over a three-dimension grid can be rather time-consuming. It is of practical importance to know whether it is necessary to tune all three parameters or whether it is possible to, say, set $\gamma = 3$ and retain robust performance over a range of possible scenarios. To investigate this question, we organize the section as follows: Section 5.1 fixes both α and γ, so that the Mnet estimator has only a single tuning parameter, similar to the LASSO. Section 5.2 tunes α, so that Mnet (and Enet) carry out tuning parameter selection in two dimensions. Finally, Section 5.3 tunes α and γ, so that Mnet tunes all three of its parameters. We compare the proposed Mnet estimator with the elastic net, LASSO, and MCP. The `ncvreg` package was used to fit all models.

The basic design of our simulations in these sections is as follows: all covariates $\{x_j\}$ marginally follow standard Gaussian distributions, but are correlated with a common correlation ρ between any two covariates. The outcome y is generated according to model (2.1), with errors drawn from the standard Gaussian distribution. For each independently generated data set, $n = 100$ and $p = 500$, with 12 nonzero coefficients equal to s and the remaining 488 coefficients equal to zero. Throughout, we consider varying the correlation ρ between 0.1 and 0.9 and the signal strength s between 0.1 and 1.5. For all methods, tuning parameters are selected on the basis of mean-squared prediction error on an independent validation data set also of size $n = 100$.

We focus primarily on the estimation accuracy of $\hat{\beta}$, as measured by mean squared error (MSE) relative to LASSO. Plots of the absolute MSE are dominated by the high-correlation and high-signal scenarios, which are not necessarily the most interesting. The LASSO, as the most widely used penalized regression method, is an obvious choice as a baseline for measuring relative estimation efficiency.

For the simulations in Sections 5.1 and 5.2, we also present results concerning variable selection accuracy, although due to space limitations, these are presented in a separate Supplementary Materials document. Comparing variable selection accuracy among methods is less straightforward than comparing estimation accuracy, due to the fact that incorrect selection are of two types: failure to select import variables and selection of unimportant variables. The LASSO, for example, tends to select more variables, both important and unimportant, than MCP and Mnet. Overall, however, the variable selection results are generally in accordance with the estimation accuracy results in terms of which methods perform well in which settings.

Zou and Hastie (2005) advocated a rescaling factor to be applied to the elastic net in order to decrease bias. We investigated the rescaling issue for the Enet and Mnet estimators, but found it to make little difference in terms of the MSE. Scaling up the estimator does decrease bias, but increases variance with little to no net benefit for either the Mnet or Enet estimators. As we found the difference between the original and rescaled estimators to be minuscule compared with the difference between Mnet and Enet themselves, we focus only on the original estimators here, as defined in Section 2.

5.1. Fixed α and γ

In this section, we fix the α parameter for Mnet at various values in the set $\{0.5, 0.7, 0.9\}$ and the γ parameter at $\gamma = 3$. By fixing both tuning parameters, the Mnet estimator, like LASSO, has just a single parameter τ to select. We compare these fixed-α versions of Mnet with the LASSO and present the results in Figure 1. For both methods, τ was selected using external validation.

Figure 1 illustrates the fact that no single value of α can ensure robust performance of the Mnet estimator over a variety of signal and correlation strengths. For example, when $\rho = 0.3$ and $s = 1.5$, Mnet ($\alpha = 0.9$) is far more accurate than the LASSO, which in turn is quite a bit more accurate than Mnet ($\alpha = 0.5$). However, when $\rho = 0.7$ and $s = 0.8$, the rankings are reversed: Mnet ($\alpha = 0.5$) is more accurate than LASSO, and LASSO is more accurate than Mnet ($\alpha = 0.9$).

The primary message then is that any fixed-α Mnet estimator is vulnerable to poor performance in certain scenarios. When estimation is relatively easy (high signal strength, low correlation), Mnet estimators with substantial weight on the L_2 penalty suffer from shrinkage-induced bias and have much higher MSE than estimators which place most of their weight on the MCP portion of the penalty. Conversely, when estimation is relatively difficult (low signal strength, high correlation), Mnet estimators with α near 1 suffer from large variance relative to

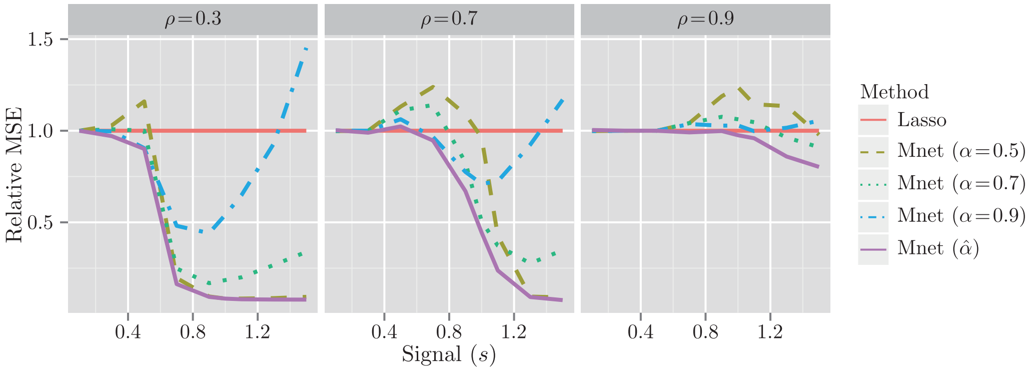

Figure 1. Relative (to the LASSO) MSE for the variable-α Mnet (with α selected by external validation) and various fixed-α Mnet estimators. MSE was calculated for each method on 100 independently generated data sets; the relative median MSEs at each point are displayed.

estimators that provide additional L_2 shrinkage. Given this clear dependence of the optimal α value on correlation, and the considerable difficulty of even estimating the correlation matrix of the coefficients in high dimensions, let alone using that estimate to determine an optimal α, empirical tuning of α seems to be warranted.

The Mnet ($\hat{\alpha}$) line in Figure 1 depicts the performance of a variable-α Mnet estimator in which the α has been tuned over a grid of reasonable values (see Section 5.2 for details). By allowing α to vary, we avoid the vulnerabilities of the fixed-α Mnet estimators, as certain α values are rarely chosen in scenarios where they perform poorly. For example, with $\rho = 0.3$ and $s = 1.1$, $\alpha = 0.9$ or $\alpha = 1$ was selected in 98% of the simulations. When $\rho = 0.7$ and $s = 0.7$, $\alpha = 0.9$ or $\alpha = 1$ was selected in only 10% of the simulations, with $\alpha = 0.5$ being the most commonly selected value, chosen in 42% of simulations. We explore variable-α Mnet estimators further in Section 5.2.

5.2. Select α, fixed γ

In this section, we compare variable-α versions of the Mnet and elastic net estimators with MCP and LASSO. For the Mnet and Enet, the α parameter was allowed to vary among $\{0.1, 0.3, \ldots, 0.9\}$, with the optimal value selected by external validation.

A comparison of the LASSO, MCP, Enet, and Mnet estimators is given in Figure 2. Here we see that the variable-α Mnet is able to achieve what the fixed-α Mnet cannot; namely, robust accuracy over the full spectrum of correlation and signal strengths. The Mnet estimates are the best or virtually identical to the best in all situations, sometimes dramatically so. In particular, for the medium

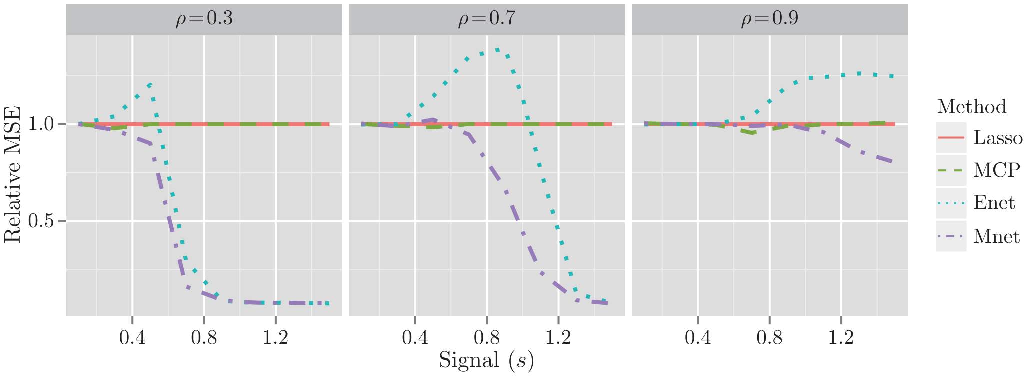

Figure 2. Relative (to the LASSO) MSE for the MCP, elastic net (Enet) and Mnet estimator. MSE was calculated for each method on 100 independently generated data sets; the relative median MSEs at each point are displayed.

correlation ($\rho = 0.7$), medium signal ($s = 1$) case, Mnet is more than twice as efficient as the other three methods.

Indeed, as was seen in Figure 1, the variable-α Mnet performs roughly as well as the best fixed-α Mnet estimator in all scenarios, indicating that little performance is lost in the model selection process of estimating the optimal value of α. There is no scenario in Figures 1 or 2 in which the variable-α Mnet estimator does poorly; notably, it always attains an estimation accuracy as good or better than that of the lasso.

In addition, Figure 2 illustrates that the addition of an L_2 penalty has a far bigger impact on MCP than on LASSO. Indeed, although there are small benefits of the elastic net over LASSO to be seen in each panel, these differences are minute compared with the differences between MCP and Mnet, as well as the differences between Mnet and Enet.

5.3. Select α and γ

Here we tune the τ, α, and γ parameters for the Mnet estimator, again investigating performance relative to LASSO and Enet. The γ parameter for Mnet was allowed to vary among $\{2, 4, 8, 16, 32\}$, with the optimal value selected by external validation. The results of the simulations are shown in Figure 3.

In contrast to Figure 2, Figure 3 indicates little benefit to tuning both α and γ; the performance of Mnet $(\hat{\alpha})$ and Mnet $(\hat{\alpha}, \hat{\gamma})$ are quite similar. Overall, the simulation results suggest that tuning γ is likely not worth the additional time and effort required.

The reason for this is likely that, to some extent, α and γ play similar roles in terms of shrinkage and balancing the bias-variance trade-off. If signal is small and heavy shrinkage is desirable, we can achieve that by increasing γ, but we

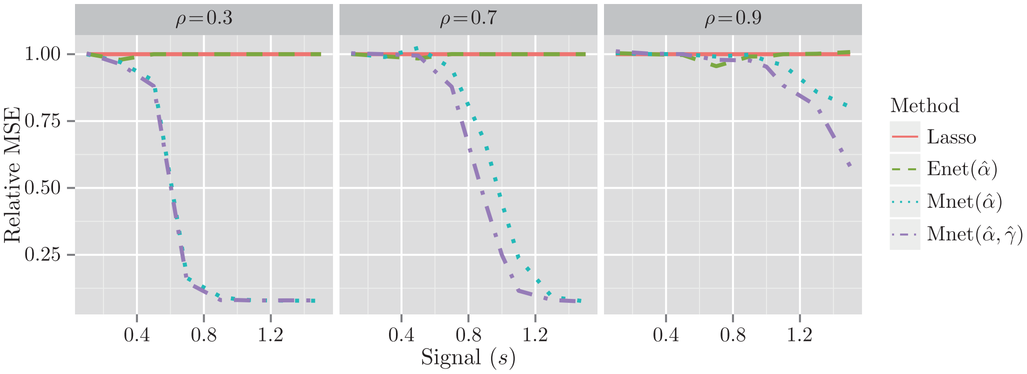

Figure 3. Relative (to the LASSO) MSE for the elastic net (Enet) and Mnet estimators. For Mnet $(\hat{\alpha}, \hat{\gamma})$, γ was selected by external validation; for Mnet $(\hat{\alpha})$, γ was fixed at 3. MSE was calculated for each method on 250 independently generated data sets; the relative median MSEs at each point are displayed.

can also increase shrinkage by lowering α, which puts more weight on the L_2 penalty. Thus, although there are scenarios where MCP does poorly relative to LASSO, and therefore where we would benefit from tuning γ, we can achieve the same benefits, if not more, by tuning α.

Figure 2 provides a nice illustration of this. For $s = 0.7$ and $\rho = 0.9$, we see that LASSO performs considerably better than MCP. We would therefore benefit from tuning γ (increasing γ to make the MCP estimates more LASSO-like). However, the Mnet results show that we obtain even better performance by fixing $\gamma = 3$ and tuning α. Specifically, at this setting, the MSE of MCP is 39% higher than that of LASSO, while the MSE of Mnet is 33% lower than that of LASSO.

Thus, α is not necessarily intrinsically more important than γ; it is more the case that two-dimensional tuning is not substantially better than fixing one and tuning the other. Here α is somewhat easier to tune since it falls within a finite range $[0, 1]$, whereas the range of γ is $[1, \infty)$, with $\gamma \approx 1$ resulting in instability and multiple local minima (Breheny and Huang (2011)).

5.4. Grouping

Sections 5.1−5.3 describe the basic properties of the Mnet estimator in terms of its performance relative to MCP, LASSO, and elastic net. In this section, we report on a small simulation study to illustrate the consequences of the grouping phenomenon described in Section 2.3.

The covariates were standard normal marginally, but their correlation structure was now block-diagonal. Specifically, the $p = 1,000$ covariates consisted of 100 blocks; each block consisted of 10 covariates, all of which shared a common

Table 1. Grouping simulation: Median values over 100 replications.

				Variables		Blocks	
	MSE	**MSPE**	**wSD**	**Selected**	**True**	**Selected**	**True**
				Concentrated			
MCP	0.0190	3.12	0.43	7	4	5	2
Mnet	0.0022	1.47	0.21	18	16	3	2
				Scattered			
MCP	0.0069	3.87	0.10	25	4	25	16
Mnet	0.0064	3.15	0.08	36	5	36	19

within-block pairwise correlation of 0.9. The blocks themselves were independent. The β coefficients for block one were all equal to 0.5; the coefficients for block two are all equal to -0.5; the coefficients in the other 98 blocks are all zero. We refer to this setting as "concentrated", since the nonzero coefficients are concentrated with respect to the blocks. We also considered a "scattered" setting, in which 20 blocks had one zero coefficient and 9 nonzero coefficients; the coefficients in the other 80 blocks were all equal to zero. The distribution of the covariates as well as the values of the regression coefficients were the same in the two settings; the only difference was the arrangement of the nonzero coefficients with respect to the correlation structure.

Table 1 presents the results of this simulation for MCP and Mnet ($\alpha = 0.5$). Here we consider two basic measures of model accuracy: mean squared estimation error (MSE) and mean squared prediction error (MSPE). We also report the number of variables and blocks selected by each method and of those, the number that were truly nonzero in the generating model. Finally, as a measure of the grouping effect, we calculated the within-group standard deviation of $\hat{\beta}$ among the nonzero groups (wSD).

As the table shows, the advantages of Mnet over MCP are amplified when the coefficient values reflect the correlation structure. In particular, the MSE of MCP is over 8 times larger than that of Mnet when the nonzero coefficients are concentrated in blocks. Furthermore, the selection properties of Mnet are desirable here: although Mnet selects a larger number of variables, it actually selects fewer blocks, concentrating those selections into correlated groups and thereby detecting far more individual nonzero coefficients.

Even when there is no relationship between the coefficient values and the correlation structure, there are still advantages to using Mnet, although the gains in efficiency are not nearly as dramatic. In addition to all the results from Sections 5.1−5.3, we can also see evidence for this in the "Scattered" section of Table 1. Mnet still yields more accurate estimates, but the improvement is only 8%. In this setting, both Mnet and MCP performed reasonably well at

Table 2. Cross-validated R^2 for Scheetz data.

	$\alpha = 1$	$\alpha = 0.75$	$\alpha = 0.5$	$\alpha = 0.25$
Enet	0.49	0.50	0.51	0.52
Mnet	0.54	0.47	0.60	0.57

detecting the nonzero blocks, but the methods struggled to select the single correct predictor from the block of highly correlated choices.

6. Application to Gene Expression Data

We use the data set reported in Scheetz et al. (2006) to illustrate the application of the proposed method in high-dimensional settings. For this data set, 120 twelve-week-old male rats were selected for eye tissue harvesting and microarray analysis. The microarrays used to analyze eye tissue RNA from these animals contain over 31,042 different probe sets (Affymetric GeneChip Rat Genome 230 2.0 Array). The intensity values were normalized using the robust multi-chip averaging method (Irizarry et al. (2003)) to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale.

The goal of the analysis is to detect genes whose expression patterns exhibit a reasonable degree of variability and which are most correlated with that of gene TRIM32. This gene has been found to cause Bardet-Biedl syndrome (Chiang et al. (2006)), a genetically heterogeneous disease of multiple organ systems including the retina. We apply regularized linear regression using the Enet/Mnet penalties, with TRIM32 expression as the response and restricted our attention to the 5,000 genes with the largest variances in expression (on the log scale). Thus, this data set had $n = 120$ and $p = 5,000$. Ten-fold cross-validation was used to select τ and α, with γ for the Mnet fixed at 3.

To compare the predictive ability of various Enet and Mnet estimators, we examine the cross-validated prediction error:

$$PE = \frac{\sum_i (y_i - \hat{y}_{(-i)})^2}{n},$$

where, for each observation i, $\hat{y}_{(-i)}$ is the estimated value of $E(Y_i|x_{i1}, \ldots, x_{ip})$ based on the estimates $\hat{\beta}$ from the model fit based on the cross-validation fraction of the data leaving out observation i. Table 2 presents the prediction error of the Enet and Mnet methods, presented as the proportion of variance explained by the model: $R^2 = PE/\widehat{\text{Var}}(y)$, where $\widehat{\text{Var}}(y)$ denotes the sample variance of the response.

Table 2 again shows that tuning α has a much larger effect for the Mnet estimators than the elastic net estimators. Here the best Mnet model outperforms the best Enet model, with Mnet ($\alpha = 0.5$) explaining 60% of the variability in

Table 3. Genes estimated to have nonzero effect using the Enet and Mnet approaches.

Probe ID	**Gene**	**Enet**	**Mnet**
1382452_at	Sdpr	0.01	
1383110_at	Klhl24	0.01	
1383749_at	Phospho1	-0.01	
1383996_at	Med26	0.03	
1376267_at		0.04	
1382673_at		0.04	0.10
1393736_at		0.01	
1375872_at		0.02	
1376747_at		0.09	0.24
1390539_at		0.04	
1379835_at		-0.02	
1379600_at		-0.01	
1387929_at	Pmf31	0.01	
1379094_at		0.01	
1386358_at		-0.01	
1384860_at	Zfp84	0.01	
1381902_at	Zfp292	0.06	0.16
1377792_at		0.03	
1378485_at		-0.01	
1378847_x_at	H2afx	-0.01	
1375577_at		-0.01	-0.03
1395888_at			-0.01

TRIM32 expression and Enet ($\alpha = 0.25$) explaining 52%. It is *not* the case that Mnet is always superior to Enet regardless of α: an analyst who decided on $\alpha = 0.75$ *a priori*, not an unreasonable choice, would only be able to explain 47% of the variability in the response – a worse performance than LASSO. Tuning α is essential for obtaining the best performance from Mnet.

In addition to superior prediction accuracy, Mnet also identifies a considerably more sparse model. Mnet ($\alpha = 0.5$) is able to explain 60% of TRIM32 using only 5 genes, compared to 21 genes selected by elastic net (including four of the five Mnet genes). The genes selected by each method appear in Table 3 along with their corresponding coefficient estimates. In addition to the sparser Mnet estimates, the other salient feature in the table is that the sizes of the Mnet estimates are much larger than their corresponding Enet estimates. In particular, both Enet and Mnet estimate that the transcript 1376747_at is most closely associated with TRIM32 expression, but the size of the Mnet coefficient is almost three times larger than the Enet coefficient. The heavy bias towards zero displayed by the elastic net cannot be meaningfully alleviated by Zou and Hastie (2005)'s rescaling proposal: the rescaling adjustment here is only 4.6%, and the rescaled estimate is still 0.09 when rounded to the nearest hundredth.

The bias reduction achieved by the Mnet in comparison with the elastic net is particularly relevant in practice, as an important goal of studies like this one is to estimate the effects of the most important genes, here 1376747_at and Zfp292. The elastic net systematically underestimates the contribution of such genes with respect to Mnet. In addition, the parsimony of the Mnet models is of considerable practical importance, not only for ease of interpretation but also to reduce the time and cost of confirmatory follow-up studies.

7. Discussion

Although we have focused on linear regression, the Mnet approach can be extended in a straightforward manner to the regression problems

$$\frac{1}{2n}\sum_{i=1}^{n}\ell(y_i,\beta_0+\sum_{j}x_{ij}\beta_j)+\sum_{j=1}^{p}\rho(|\beta_j|;\lambda_1,\gamma)+\frac{1}{2}\lambda_2\|\beta\|^2,$$

where ℓ is a given loss function. This formulation includes generalized linear models, censored regression models and robust regression. For instance, for generalized linear models such as logistic regression, we take ℓ to be the negative log-likelihood function. For Cox regression, we take the empirical loss function to be the negative partial likelihood. For loss functions other than least squares, further work is needed to study the computational algorithms and theoretical properties of the Mnet estimators, although we note that the `ncvreg` package has already been extended to fit Mnet-penalized logistic regression and Cox regression models.

Our results provide insight into the strengths and weaknesses of MCP and how minimax concave penalized regression can be stabilized through the incorporation of an addition L_2 penalty to produce a new procedure similar in spirit to the elastic net. Our simulation results show that MCP alone often fails to outperform the LASSO in the presence of correlated features but that, provided one selects the α parameter empirically, the proposed Mnet estimator achieves robust performance gains over the LASSO over a wide range of settings. Furthermore, our theoretical results show that Mnet has the oracle selection property under reasonable conditions.

Although in principle, the proposed method has three tuning parameters, our simulation results indicate that there is little benefit in tuning both γ and α. Efficient estimation across a wide range of settings is achieved by fixing γ (for example $\gamma = 2.7$ as suggested in Zhang (2010), for models with standardized predictors) and selecting only τ and α. Following this approach, the selection of tuning parameters with Mnet is no more complicated than that of the elastic net.

The prediction performance of Mnet is typically similar to that of the elastic net. The main advantage of Mnet over Enet is that it achieves this prediction performance using a smaller set of features. This is advantageous for several reasons, including a lower false discovery rate and a lower cost of follow-up analyses and assays studying the selected features, in addition to the more straightforward benefit of obtaining a simpler and more parsimonious model.

Supplementary Materials

Additional results concerning variable selection accuracy from the simulation studies of Section 5, as well as proofs of Proposition 1 and Theorems 1 and 2, are included in the Supplementary Materials.

Acknowledgements

The authors wish to thank two anonymous referees and an associate editor for their helpful comments. The research of Huang is partially supported by NSF grant DMS-1208225. The research of Ma is partially supported by NIH grant R01CA142774. The research of Zhang is partially supported by NSF Grants DMS-1129626, DMS-1209014 and DMS-1513378 and NSA Grant H98230-15-1-0040.

References

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression methods. *Ann. Appl. Statist.* **5**, 232-253.

Chiang, A., Beck, J., Yen, H.-J., Tayeh, M., Scheetz, T., Swiderski, R., Nishimura, D., Braun, T., Kim, K.-Y., Huang, J., Elbedour, K., Carmi, R., Slusarski, D., Casavant, T., Stone, E. and Sheffield, V. (2006). Homozygosity mapping with SNP arrays identifies a novel gene for Bardet-Biedl Syndrome (BBS10). *Proc. Natl. Acad. Sci.* **103**, 6287-6292.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

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

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **35**, 302-332.

Fu, W. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.* **7**, 397-416.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264.

Jia, J. and Yu, B. (2010). On model selection consistency of the elastic net when $p \gg n$. *Statist. Sinica* **20**, 595-611.

Mazumder, R., Friedman, J. and Hastie, T. (2011). *SparseNet*: Coordinate descent with nonconvex penalties. *J. Amer. Statist. Assoc.* **106**, 1125-1138.

Scheetz, T., Kim, K.-Y., Swiderski, R. E., Philp1, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., Sheffield, V., and Stone, E. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proc. Natl. Acad. Sci.* **103**, 14429-14434.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wu, T. and Lange, K. (2008). Coordinate descent procedures for lasso penalized regression. *Ann. Appl. Statist.* **2** 224-244.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *J. Roy. Statist. Soc. Ser. B* **69**, 143-161.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Machine Learning Res.* **7**, 2541-2567.

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

Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a divergent number of parameters. *Ann. Statist.* **37**, 1733-1751.

Department of Statistics & Actuarial Science, University of Iowa, Iowa City, Iowa 52242, USA.

E-mail: jian-huang@uiowa.edu

Department of Biostatistics, University of Iowa, Iowa City, Iowa 52242, USA.

E-mail: patrick-breheny@uiowa.edu

Quantitative Biomedical Research Center, UT Southwestern Medical Center,Dallas, TX 75390, USA.

E-mail: sanginlee44@gmail.com

School of Public Health, Yale University, New Haven, CT 06520, USA.

E-mail: shuangge@gmail.com

Department of Statistics and Biostatistics, Rutgers University, New Brunswick, NJ 08901, USA.

E-mail: czhang@stat.rutgers.edu

(Received August 2014; accepted September 2015)