The Mnet Method for Variable Selection

Jian Huang[1], Patrick Breheny[2], Shuangge Ma[3] and Cun-Hui Zhang[4]
[1]University of Iowa, [2]University of Kentucky, [3]Yale University and [4]Rutgers University

May 2010
The University of Iowa
Department of Statistics and Actuarial Science
Technical Report No. 402

The Mnet Method for Variable Selection

Jian Huang[1], Patrick Breheny[2], Shuangge Ma[3] and Cun-Hui Zhang[4]
[1]University of Iowa, [2]University of Kentucky, [3]Yale University and [4]Rutgers University

Summary. We propose a new penalized approach for variable selection using a combination of minimax concave and ridge penalties. The proposed method is designed to deal with $p \geq n$ problems with highly correlated predictors. We call the propose approach the Mnet method. Similar to the elastic net of Zou and Hastie (2005), the Mnet also 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 apply the coordinate descent algorithm to compute the Mnet estimates. Simulation studies show that the Mnet has better performance in variable selection in the presence of highly correlated predictors than the elastic net method. An example is used to illustrate the application of the Mnet method.

Some key words. 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, which include estimators based on the bridge penalty (Frank and Friedman 1993), the ℓ_1 penalty or the least absolute shrinkage and selection operator (LASSO, Tibshirani 1996; Chen, Donoho and Saunders 1998), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), and the minimum concave penalty (MCP, Zhang 2010). These methods provide a computationally feasible way for variable selection in high-dimensional settings. Much progress has been made in understanding the theoretical properties of these methods.

While these methods have many attractive properties, they also 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 only 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 proposed the elastic net (Enet) method, 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 are 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 $p < n$.

Therefore, 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 new penalized approach that uses a combination of the MCP and ℓ_2 penalty. We call this new method 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. In addition, because the Mnet uses the MCP instead of the ℓ_1 penalty for selection, it has two important advantages. First, the Mnet is selection consistent under a sparse Riezs condition on the 'ridge design matrix', which only requires a submatrix of this matrix to be nonsingular. This condition is different from the irrepresentable condition and is usually less restrictive, especially in high-dimensional settings (Zhang, 2010). Second, 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 estimate the selected coefficients using ridge regression. The Enet does not have such an oracle property because the shrinkage introduced by the ℓ_1 penalty results in nonnegligible bias for large coefficients.

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 Mnet 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. All the technical proofs are provided in the Appendix.

2 The Mnet estimator

Consider a 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 j regression coefficient and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ is the vector of random errors. 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

To define the Mnet estimator, we first provide a brief description of the MCP introduced by Zhang (2010). The MCP is defined as

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

where λ_1 is a penalty parameter and γ is a regularization parameter. Here x_+ is the non-negative part of x, i.e., $x_+ = x1_{\{x \geq 0\}}$. The MCP can be easily understood by considering its derivative, which is

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

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 $\gamma \to 0+$.

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

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

We note that the Enet criterion uses the ℓ_1 penalty in the first penalty term. In contrast, 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 because the it imposes the same shrinkage on small and large coefficients. 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 the 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 et al. (2009).

2.2 Orthonormal designs

To gain some insights into the characteristics of the Mnet estimator, we look at the case when 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 $z = n^{-1}\sum_{i=1} x_{ij}y_i$ 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 $\tilde{\theta}_{Mnet}$ of (2.6) is

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

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 Enet. The naive Enet (nEnet) estimator is

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

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 removes the ridge shrinkage factor by multiplying the naive Enet by $(1+\lambda_2)$ to obtain he Enet estimator

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

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, which can be written as

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

which 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. Also, since here we focus on the variable selection property of the Mnet and rescaling does not affect selection results, we will not consider rescaling in this article.

2.3 Grouping effect

Similar to the Enet, the Mnet also has the grouping effect. It tends to select or drop strongly correlated predictors together. This grouping property is due to the ℓ_2 penalty term. The following proposition describes this property.

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 $\lambda_2 > 0$. Denote*

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

For $\rho_{jk} \geq 0$, we have

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

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

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

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 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 criterions with convex penalties such as LASSO (Fu 1998; Friedman et al. 2007; Wu and Lange 2007). It has been proposed to calculate the MCP estimates (Breheny and Huang 2009). This 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 problem, then, is to minimize M with respect to β_j, given current values for the regression coefficients $\tilde{\beta}_k$. Define

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

Denote $\tilde{y}_{ij} = \sum_{k\neq j} x_{ik}\tilde{\beta}_k, \tilde{r}_{ij} = y_i - \tilde{y}_{ij},$ 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

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

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

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

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:

(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 $\tilde{y}_i = \sum_{j=1}^{n} x_{ij}\tilde{\beta}_j^{(s)}$ 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 $O(2n)$ 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 following reparameterization: $\tau = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1/\tau$. Using this parametrization, 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 may 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 both 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. The conditions required for this to happen are established in the proposition below.

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 β on $\mathbb{R}^p$ if and only if $\gamma > 1/(c_{\min} + \lambda_2)$.*

The above proposition 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 may 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 $\beta_j \neq 0$. The issue of local convexity is explored in greater detail in Breheny and Huang (2009).

4 Selection properties

In this section, we study the selection properties of the Mnet estimator $\hat{\beta}_{Mnet}$ in (2.5). We provide sufficient conditions under which the Mnet estimator is sign consistent and equals the oracle ridge estimator defined in (4.1) below.

For simplicity of notation, we write $\hat{\beta} = \hat{\beta}_{Mnet}$. Denote $\Sigma = n^{-1}X'X$. For any $A \subseteq \{1, \ldots, p\}$, define

$$X_A = (x_j, j \in A), \quad \Sigma_A = \frac{1}{n}X_A'X_A.$$

Let the true value of the regression coefficient be $\beta^o = (\beta_1^o, \ldots, \beta_p^o)'$. Denote $O = \{j : \beta_j^o \neq 0\}$, which is the oracle set of indices of the predictors with nonzero coefficients in the underlying model. Let $\beta_* = \min\{|\beta_j|, j \in \mathcal{O}\}$ and set $\beta_* = \infty$ if $\mathcal{O}$ is empty, that is, if all the regression coefficients are zero. Denote the cardinality of $\mathcal{O}$ by $|\mathcal{O}|$ and let $d^o = |\mathcal{O}|$. So d^o is the number of nonzero coefficients. Define

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

This is the oracle ridge estimator. Of course, it is not a real estimator, since 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.

Denote

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

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

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

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

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

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

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

We note that the upper bound on the probability of selection error is nonasymptotic. (A1a) is standard in linear regression. (A1b) is concerned with the tail probabilities of the error distribution. Here we allow non-normal and heavy tail error distributions. 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 that 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.

The following corollary is an immediate consequence of Theorem 1.

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*

$$\mathrm{P}(sgn(\hat{\beta}) \neq sgn(\beta^o) \text{ or } \hat{\beta} \neq \hat{\beta}^o) \to 0.$$

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 any 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 the 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 $A \subseteq \{1, \ldots, p\}$, define

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

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

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

Here 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.

Recall d^o is the number of nonzero coefficients. In addition to (A1), we also need the following condition.

(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 $K_* = (c^* + \lambda_2)/(c_* + \lambda_2) - (1/2)$.

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

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

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). Define

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

Theorem 2 *Suppose that (A1) and (A2) hold. Also, suppose that*

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

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

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

Theorem 2 has the following corollary.

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*

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

Theorem 2 and Corollary 2 provide sufficient conditions for sign consistency and 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 Appendix, 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, 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, consider the simplest case when the error distribution has sub-gaussian tails ($\alpha = 2$) and $\sqrt{c^*}/(m\sqrt{n}(c_* + \lambda_2)^2) \leq 1$ in (4.7) for sufficiently large n, then we can have $p - d^o = \exp(o(n))$, where $o(n)/n \to 0$.

5 Numerical studies

5.1 Penalty parameter selection

For the Mnet estimator parameterized according to (τ, α, γ) described in Section 3.2, there are two tuning parameters, γ and α, in addition to the parameter τ, which controls the overall degree of regularization. As $\alpha \to 1$, the Mnet becomes MCP; as $\alpha \to 0$, it becomes equivalent to ridge regression. Large values of α and small values of γ tend to produce more

sparse models; however, as is clear from Proposition 2, they are also more likely to produce a nonconvex objective function. Proper choices for α and γ will thus depend on a number of factors such as the relative sizes of n and p, the sparsity and signal-to-noise ratio of the underlying data-generating process, and the multicollinearity of the covariates. Several data driven procedures are available for tuning parameter selection. Here, as in Zou and Hastie (2005), we use ten-fold cross validation to select tuning parameters τ, α, and, for Mnet, γ. For both Mnet and Enet, there were 100 candidate values of τ and four candidates for α: 1, 0.9, 0.5, and 0.1. For Mnet, the candidate values for γ were 2.5 and 6. We found that the Mnet is not sensitive to small changes in γ and that selecting γ from the two candidate values ($\gamma = 2.5$ or 6) worked well.

5.2 Simulation studies

Our simulation studies examine the performance of the Mnet estimator in comparison with the Enet in two distinct settings: one in which all covariates are uniformly correlated with each other, and another in which correlation is present within small groups of covariates. In the simulations, the magnitude of the regression coefficients β, the number of nonzero coefficients p_1 , and the correlation ρ were varied and their impact on the estimation, prediction, and variable selection properties of the two methods were investigated.

5.2.1 Uniform correlation

Covariates were randomly generated from the multivariate normal distribution with zero mean and correlation matrix having 1 along the diagonal and correlation coefficient ρ at all other entries. The response y was generated according to (2.1) with standard normal errors. For each independently generated data set, $n = p = 100$. In the generating model, p_1 of the variables had nonzero coefficient β, while the rest were set to zero.

Estimation accuracy, measured by mean squared error (MSE) is plotted in Figure 1. The dominant trend depicted by the figure is the improvement in accuracy of Mnet relative to Enet for large values of β. This trend should not come as a surprise, since the purpose of the MCP component of the Mnet penalty is to eliminate the downward bias of the lasso for large coefficients. Note, however, that the downward bias reduces variance and is capable of improving estimation for small model coefficients. This general trend is seen for all combinations of p_1 of ρ. However, the trend is weakest in the presence of high correlation. In such settings, cross-validation selects small values of α for both Mnet and Enet, leading both methods to produce estimators similar to each other and to ridge regression.

The variable selection properties of Mnet and Enet, as measured by the false discovery rate (FDR), are plotted in Figure 2. Because it lacks the built-in ability to relax downward

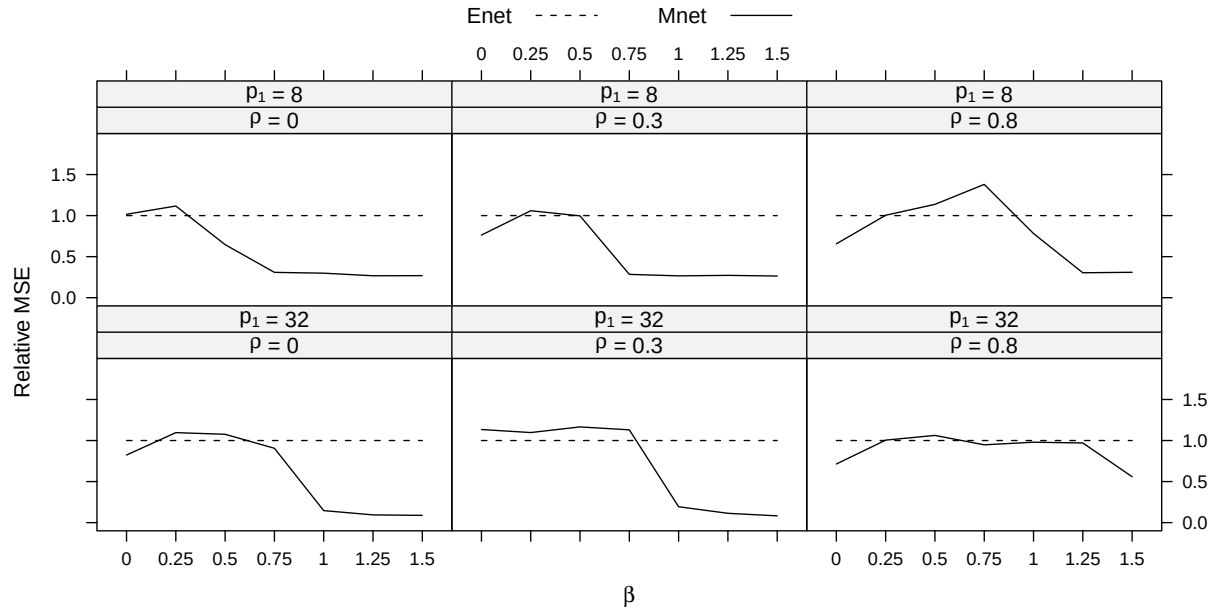

Figure 1: Relative (to the elastic net) mean squared error for the Mnet estimator in the uniform correlation simulation of Section 5.2.1. MSE was calculated for each method on 250 independently generated data sets; the relative median MSEs at each point are displayed.

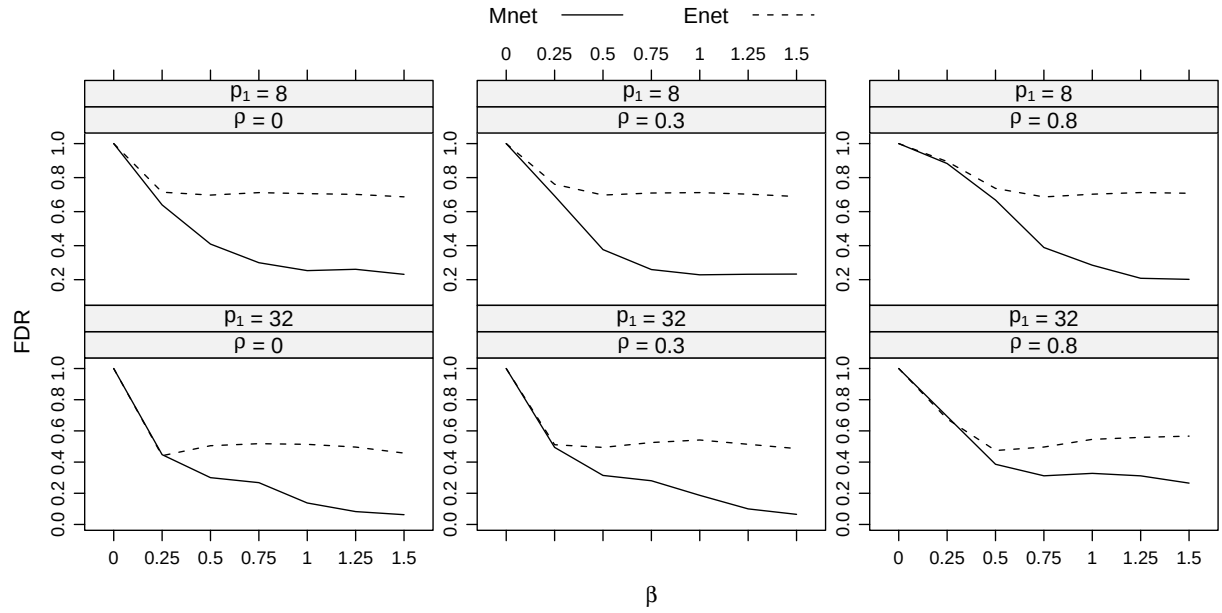

Figure 2: False discovery rates for the Mnet and Enet estimators in the uniform correlation simulation of Section 5.2.1. FDR is calculated for each method on the same independently generated 250 data sets as 1.

bias that Mnet possesses, the elastic net must select lower values of λ to reduce downward bias. Doing so, however, not only reduces bias but allows additional variables to enter the model. This results in a slight increase in the probability that covariate with a nonzero coefficient is selected, but a much larger increase in the probability that covariates with zero coefficients will be selected. The FDR – the proportion of selected variables that have a coefficient equal to zero in the underlying model – measures the overall success of the method at selecting variables that are truly related to the outcome. Mnet has a lower FDR than Enet for all values of β, p_1, and ρ depicted in Figure 2. The difference is minor for small β, but quite drastic for coefficients with large regression coefficients.

The results in this section and the next compare the unscaled versions of Mnet and Enet. Scaled versions of the two estimators were also investigated, but omitted from the preceding plots for the sake of clarity. In general, the unscaled version of Mnet outperformed the scaled version, while the scaled version of Enet outperformed the unscaled version. The differences, however, were negligible in comparison with the differences between Mnet and Enet.

5.2.2 Grouped correlation

For the simulations in this section, a grouping structure was built into the covariates as follows. For each covariate x_{ij} in group g, $x_{ij} = a_j z_{ig} + \epsilon_{ij}$, where z_{ig} and ϵ_{ij} both follow a standard normal and a_j can be adjusted as desired to vary the level of within-group correlation. In our simulations, we used a group size of three. In addition to the grouped covariates with positive model coefficients, independent covariates with zero coefficients were also generated from the standard normal distribution. This produces a design matrix in which each group consists of three correlated covariates with nonzero coefficients and with no correlation between groups or between the groups and the covariates with zero coefficients (*i.e.,* the covariance matrix is block diagonal for the covariates with nonzero coefficients and diagonal elsewhere).

In addition to specifying constant values of the within-group correlation, more realistic mixed settings were also constructed. In the mixed correlation setting, a_j was generated from the exponential distribution with rate 1. This produces pairwise correlations among group members that can range from 0 to 1, with a mean correlation of about 0.3. In this setting, the correlation varies from data set to data set, but remain constant from observation to observation within a data set.

A full series of simulations similar to those in Section 5.2.1 was conducted. However, only the results for $p_1 = 6$ and mixed correlation are presented in Figure 3; the results for other values of p_1 and ρ are similar. Figure 3 displays the effect of changing the size of the regression coefficient β upon estimation accuracy as measured by MSE, prediction accuracy

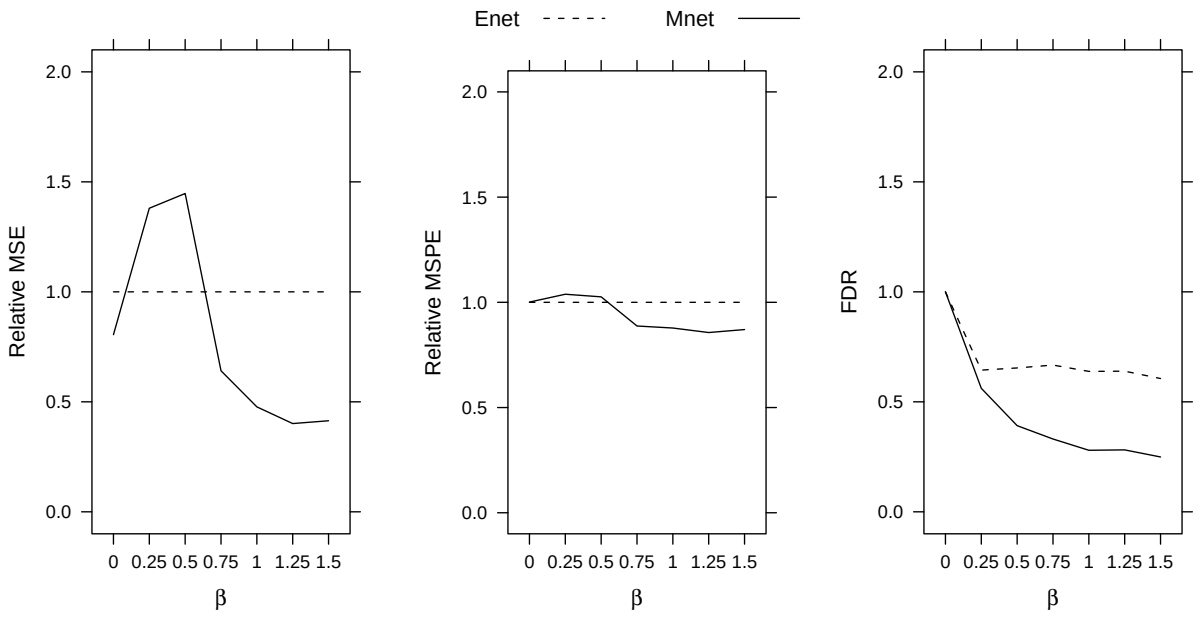

Figure 3: Relative mean squared error, relative mean squared prediction error, and false discovery rate for the Mnet and Enet estimators. All results are based on the same independently generated 250 data sets. Relative MSE and MSPE are calculated relative to Enet and are based on the median of the 250 replications.

as measured by mean squared prediction error (MSPE), and variable selection as measured by FDR.

The trends previously remarked upon with reference to Figures 1 and 2 and present as well in Figure 3. However, Figure 3 also illustrates the tradeoffs inherent in regression modeling with correlated predictors. In comparison to Mnet, the models produced by Enet include a large number of predictors that have been heavily shrunken towards zero; this results in a high FDR, improved estimation for small regression coefficients, and poorer estimation of large regression coefficients. However, the differences between Mnet and Enet with respect to prediction are much smaller.

When the number of coefficients is large and multicolinearity is present, several models may fit the data equally well despite large differences in their underling structure – this is referred to as "model multiplicity" in Breiman (2001). In such cases, there is insufficient information present in the data to guide the selection of one model versus another. Mnet produces more sparse models, but there is no way of knowing whether this reflects the underlying reality based on the data alone. In practice, scientific knowledge and research goals may provide this guidance. It is worth mentioning, however, that the prediction accuracy of Mnet and Enet are not always similar, even in the presence of correlation. For example, with

a uniform (*i.e.*, not grouped) correlation of 0.3, $p_1 = 32$ nonzero coefficients, and $\beta = 1.5$, the prediction error of Enet is five times larger than that of Mnet.

5.3 Rat eye 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 tissue harvesting from the eyes and for microarray analysis. The microarrays used to analyze the RNA from the eyes of 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 (Irizzary et al. 2003) to obtain summary expression values for each probe set. Gene expression levels were analyzed on a logarithmic scale.

We are interested in finding the genes whose expression are most variable and correlated with that of gene TRIM32. This gene was recently found to cause Bardet-Biedl syndrome (Chiang et al. 2006), which is a genetically heterogeneous disease of multiple organ systems including the retina. One approach to finding the genes that are most related to TRIM32 is to use regression analysis. Since it is expected that the number of genes associated with gene TRIM32 is small and we are only interested in genes that are most variable, we compute the variances of gene expressions and consider the 500 genes with largest variances. We then standardize gene expressions to have zero mean and unit variance.

We apply the Enet and Mnet. For both approaches, tuning parameters are selected using ten-fold cross validation as described above. The Enet identifies 30 genes, and the Mnet identifies 26 genes. Gene information and corresponding nonzero estimates are provided in Table 1. The two sets of identified genes have 11 in common. Examination of Table 1 suggests that, for the overlapped genes, the magnitudes of estimates are in general not equal. However, they have the same signs, which suggest similar biological conclusions. We further investigate prediction performance using a 10-fold cross validation approach. The mean squared prediction errors are 1.804 for Enet and 1.737 for Mnet. Although the prediction performance of the Mnet is only slightly smaller than that of the Enet (which is consistent with findings in simulation), the Mnet selects a smaller model with a more focused set of candidate genes related to TRIM32, which makes it easier to carry out further confirmation studies.

6 Discussion

The Mnet can be applied to other regression problems, for example, in the context of the general linear models, we can use

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

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.

Our theoretical results gave insights into the characteristics of the Mnet estimator. They show that the Mnet has the oracle selection property under reasonable conditions. However, these conditions are concerned with penalty parameters that are not determined based on data. Whether the results are applicable to the case where the penalty parameters are selected using cross validation or other data driven procedures is unknown. This is an important and challenging problem that requires further investigation, but is beyond the scope of the current paper.

7 Appendix

In the Appendix, we prove Proposition 1 and Theorems 1 and 2.

Proof of Proposition 1 The jth estimated coefficient $\hat{\beta}_j$ must satisfy the KKT conditions,

$$\begin{cases}-\frac{1}{n}x_j'(y-X\hat{\beta})+\lambda_1(1-|\hat{\beta}_j|/(\gamma\lambda_1))_+\mathrm{sgn}(\hat{\beta}_j)+\lambda_2\hat{\beta}_j=0, & \hat{\beta}_j\neq 0\\ |x_j'(y-X\hat{\beta})|\leq\lambda_1, & \hat{\beta}_j=0.\end{cases}$$

Let $\hat{r}=y-X\hat{\beta}$ and $\hat{z}_j=n^{-1}x_j'\hat{r}$. After some calculation, we have, if $\gamma\lambda_2>1$,

$$\hat{\beta}_j=\begin{cases}0, & \text{if } |\hat{z}_j|\leq\lambda_1,\\ \mathrm{sgn}(\hat{z}_j)\left|\frac{\gamma(|\hat{z}_j|-\lambda_1)}{\gamma\lambda_2-1}\right|, & \text{if } \lambda_1<|\hat{z}_j|<\gamma\lambda_1\lambda_2,\\ \lambda_2^{-1}\hat{z}_j, & \text{if } |\hat{z}_j|\geq\gamma\lambda_1\lambda_2;\end{cases}$$

and if $\gamma\lambda_2\leq 1$,

$$\hat{\beta}_j=\begin{cases}0 & \text{if } |\hat{z}_j|\leq\lambda_1,\\ \lambda_2^{-1}\hat{z}_j & \text{if } |\hat{z}_j|>\lambda_1.\end{cases}$$

First, suppose that x_j and x_k are positively correlated. Based on the above expressions, we can show that

$$|\hat{\beta}_j - \hat{\beta}_k| \leq \xi |\hat{z}_j - \hat{z}_k|,$$

where ξ is given in (2.8). By the Cauchy-Schwarz inequality, $|\hat{z}_j - \hat{z}_k| = n^{-1}|(x_j - x_k)'\hat{r}| \leq n^{-1}\|x_j - x_k\|\|\hat{r}\| = n^{-1/2}\sqrt{2(1-\rho_{jk})}\|\hat{r}\|$. Since $M(\hat{\beta};\lambda) \leq M(0;\lambda)$ by the definition of $\hat{\beta}$, we have $\|\hat{r}\| \leq \|y\|$. Therefore

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

For negative ρ_{jk}, we only need to change the sign of z_k and use the same argument. □

To prove Theorems 1 and 2, we first need the lemma below. Let $\psi_\alpha(x) = \exp(x^\alpha) - 1$ for $\alpha \geq 1$. For any random variable X its ψ_α-Orlicz norm $\|X\|_{\psi_\alpha}$ is defined as $\|X\|_{\psi_\alpha} = \inf\{C > 0 : E\psi_\alpha(|X|/C) \leq 1\}$.

Lemma 1 *Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed random variables with $\mathrm{E}\varepsilon_i = 0$ and $Var(\varepsilon_i) = 1$. Furthermore, suppose that $P(|\varepsilon_i| > x) \leq K\exp(-Cx^\alpha), i = 1, \ldots, n$ for constants C and K, and $1 \leq \alpha \leq 2$. Let $c_1, \ldots, c_n$ be constants satisfying $\sum_{i=1}^n c_i^2 = 1$. Let $X = \sum_{i=1}^n c_i\varepsilon_i$.*

(i) $\|X\|_{\psi_\alpha} \leq K_\alpha\left\{1 + (1+K)^{1/\alpha}C^{-1/\alpha}\alpha_n\right\}$, where K_α is a constant only depending on α, C and K.

(ii) Let $X_1, \ldots, X_m$ be any random variables whose Orlicz norms satisfy the inequality in (i). For any $b_n > 0$,

$$\mathrm{P}\left(\max_{1\leq j\leq m}|X_j| \geq b_n\right) \leq \frac{K_1\alpha_n(\log(m+1))^{1/\alpha}}{b_n}$$

for a positive constant K_1 only depending on α, C and K.

This lemma follows from Lemma 2.2.1 and Proposition A.1.6 of Van der Vaart and Wellner (1996). We omit the proof.

Proof of Theorem 1. Since $\hat{\beta}^o$ is the oracle ridge regression estimator, we have $\hat{\beta}^o_j = 0$ for $j \notin \mathcal{O}$ and

$$-\frac{1}{n}x_j'(y - X\hat{\beta}^o) + \lambda_2\hat{\beta}^o_j = 0, \quad \forall j \in \mathcal{O}. \tag{7.1}$$

If $|\hat{\beta}^o_j| \geq \gamma\lambda_1$, then $\rho'(|\hat{\beta}^o_j|;\lambda_1) = 0$. Since $c_{\min} + \lambda_2 > 1/\gamma$, the criterion (2.4) is strictly convex. By the KKT conditions, $\hat{\beta} = \hat{\beta}^o$ and $\mathrm{sgn}(\hat{\beta}) = \mathrm{sgn}(\beta^o)$ in the intersection of the events

$$\Omega_1(\lambda) = \left\{\max_{j\notin\mathcal{O}}\left|n^{-1}x_j'(y - X\hat{\beta}^o)\right| < \lambda_1\right\} \text{ and } \Omega_2(\lambda) = \left\{\min_{j\in\mathcal{O}}\mathrm{sgn}(\beta^o_j)\hat{\beta}^o_j \geq \gamma\lambda_1\right\}. \tag{7.2}$$

We first bound $1 - \mathrm{P}(\Omega_1(\lambda))$. Let $\hat{\beta}_{\mathcal{O}} = (\hat{\beta}_j, j \in \mathcal{O})'$ and $Z = n^{-1/2}X$. Let $\Sigma_{\mathcal{O}}(\lambda_2) = \Sigma_{\mathcal{O}} + \lambda_2 I_{\mathcal{O}}$. By (7.1) and using $y = X_{\mathcal{O}}\beta^o_{\mathcal{O}} + \varepsilon$,

$$\hat{\beta}^o_{\mathcal{O}} = \frac{1}{n}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)X'_{\mathcal{O}}y = \Sigma^{-1}_{\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}}\beta^o_{\mathcal{O}} + \frac{1}{\sqrt{n}}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)Z'_{\mathcal{O}}\varepsilon. \tag{7.3}$$

Thus

$$\hat{\beta}^o_{\mathcal{O}} - \beta^o_{\mathcal{O}} = \frac{1}{\sqrt{n}}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)Z'_{\mathcal{O}}\varepsilon + \{\Sigma^{-1}_{\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}} - I_{\mathcal{O}}\}\beta^o_{\mathcal{O}}. \tag{7.4}$$

It follows that

$$\frac{1}{n}x'_j(y - X\hat{\beta}^o) = \frac{1}{n}x'_j\{I_n - Z_{\mathcal{O}}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)Z'_{\mathcal{O}}\}\varepsilon - \frac{1}{\sqrt{n}}x'_jZ_{\mathcal{O}}\{\Sigma^{-1}_{\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}} - I_{\mathcal{O}}\}\beta^o_{\mathcal{O}}.$$

Denote

$$T_{j1} = \frac{1}{n}x'_j\{I_n - Z_{\mathcal{O}}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)Z'_{\mathcal{O}}\}\varepsilon, \quad T_{j2} = -\frac{1}{\sqrt{n}}x'_jZ_{\mathcal{O}}\{\Sigma^{-1}_{\mathcal{O}}(\lambda_2)\Sigma_{\mathcal{O}} - I_{\mathcal{O}}\}\beta^o_{\mathcal{O}}.$$

First consider T_{j1}. Write $T_{j1} = n^{-1/2}\sigma\|a_j\|(a_j/\|a_j\|)'(\varepsilon/\sigma)$, where $a_j = n^{-1/2}\{I_n - Z_{\mathcal{O}}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)Z'_{\mathcal{O}}\}x_j$. Since $n^{-1/2}\|x_j\| = 1$, we have $\|a_j\| \le 1$. By Lemma 1,

$$\begin{aligned}
\mathrm{P}(\max_{j\notin\mathcal{O}}|T_{j1}| \ge \lambda_1/2) &\le \mathrm{P}(n^{-1/2}\sigma \max_{j\notin\mathcal{O}}|(a_j/\|a_j\|)'(\varepsilon/\sigma)| \ge \lambda_1/2) \\
&\le 2K_1\alpha_n\frac{\sigma\log^{1/\alpha}(p - d^o + 1)}{\sqrt{n}\lambda_1},
\end{aligned} \tag{7.5}$$

where α_n is given in (4.2).

For T_{j2}, we have $T_{j2} = n^{-1/2}\lambda_2 x'_jZ_{\mathcal{O}}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)\beta^o_{\mathcal{O}}$. Since

$$n^{-1/2}\lambda_2|x'_jZ_{\mathcal{O}}\Sigma^{-1}_{\mathcal{O}}(\lambda_2)\beta^o_{\mathcal{O}}| \le \lambda_2(c_1 + \lambda_2)^{-1}\sqrt{c_2}\|\beta^o\|,$$

we have $|T_{j2}| < \lambda_1/2$ for every j if

$$\lambda_1/2 > \lambda_2(c_1 + \lambda_2)^{-1}\sqrt{c_2}\|\beta^o\|. \tag{7.6}$$

Thus by (7.5), when (7.6) holds, $1 - \mathrm{P}(\Omega_1(\lambda)) \le \pi_1$.

Now consider the event Ω_2. Let e_j be the jth unit vector of length d^o. By (7.4),

$$\hat{\beta}^o_j - \beta^o_j = S_{j1} + S_{j2},\ j \in \mathcal{O},$$

where $S_{j1} = n^{-1}e'_j(\Sigma_{\mathcal{O}} + \lambda_2 I)^{-1}X'_{\mathcal{O}}\varepsilon$ and $S_{j2} = -\lambda_2 e'_j(\Sigma_{\mathcal{O}} + \lambda_2 I)^{-1}\beta^o_{\mathcal{O}}$. Therefore, $\mathrm{sgn}(\beta^o_j)\hat{\beta}^o_j \ge \gamma\lambda_1$ if $|\beta^o_j| + \mathrm{sgn}(\beta^o_j)(S_{j1} + S_{j2}) \ge \gamma\lambda_1$, which in turn is implied by

$$|S_{j1} + S_{j2}| \le \beta^o_* - \gamma\lambda_1,\ \forall j.$$

It follows that $1-\mathrm{P}(\Omega_2(\lambda)) \le \mathrm{P}(\max_{j\in\mathcal{O}}(|S_{j1}+S_{j2}| > \beta_*^o - \gamma\lambda_1)$. Since $|S_{j2}| \le \lambda_2\|\beta^o\|/(c_1+\lambda_2)$, we have $|S_{j2}| < (\beta_*^o - \gamma\lambda_1)/2$ if $\beta_*^o > \gamma\lambda_1 + 2\lambda_2\|\beta^o\|/(c_1+\lambda_2)$. Similarly to (7.5), by Lemma 1, when $\beta_*^o > \gamma\lambda_1 + 2\lambda_2\|\beta^o\|/(c_1+\lambda_2)$,

$$\mathrm{P}(\max_{j\in\mathcal{O}}(|S_{j1}+S_{j2}| > \beta_*^o - \gamma\lambda_1) \le 2K_1\alpha_n \frac{\sigma\sqrt{c_2}\log^{1/\alpha}(d^o+1)}{\sqrt{n}(\beta_*^o-\gamma\lambda_1)(c_1+\lambda_2)}. \tag{7.7}$$

By (7.7) and the restrictions on λ_1 and β_*^o, $1-P(\Omega_2(\lambda)) \le \pi_2$. □

Proof of Theorem 2. Let

$$\tilde{y} = \begin{pmatrix} y \\ 0_p \end{pmatrix},\ \tilde{X} = \begin{pmatrix} X \\ \sqrt{n\lambda_2}\, I_p \end{pmatrix},$$

where 0_p is a p-dimensional vector of zeros. We have

$$\hat{\beta}(\lambda) = \operatorname*{argmin}_b \{\frac{1}{2n}\|\tilde{y}-\tilde{X}b\|_2^2 + \sum_{j=1}^p \rho(|b_j|,\lambda_1)\}.$$

Thus the Mnet estimator can be considered an MCP estimator based on $(\tilde{y},\tilde{X})$.

Denote $\tilde{P}_B = \tilde{X}_B(\tilde{X}_B'\tilde{X}_B)^{-1}\tilde{X}_B'$. For $m \ge 1$ and $u \in \mathbb{R}^n$, define

$$\tilde{\zeta}(u;m,\mathcal{O},\lambda_2) = \max\left\{ \frac{\|(\tilde{P}_B - \tilde{P}_{\mathcal{O}})v\|_2}{(mn)^{1/2}} : v=(u',0_p')',\mathcal{O}\subseteq B\subseteq\{1,\dots p\},\ |B| = m+|\mathcal{O}| \right\}.$$

Here $\tilde{\zeta}$ depends on λ_2 through $\tilde{P}$. We make this dependence explicit in the notation. By Lemma 1 of Zhang (2010), in the event

$$\lambda_1 \ge 2\sqrt{c^*}\,\tilde{\zeta}(y;m,\mathcal{O},\lambda_2) \tag{7.8}$$

for $m = d^* - d^o$, we have

$$\#\{j : \hat{\beta}_j \ne 0\} \le (K_*+1)d^o \equiv p^*.$$

Thus in the event (7.8), the original p-dimensional problem reduces to a p_*-dimensional problem. Since $p_* \le d^*$, the conditions of Theorem 2 implies that the conditions of Theorem 1 are satisfied for $p = p_*$. So the result follows from Theorem 1.

Specifically, let τ_n be as in (4.2) and λ_n^* as in (4.7). Let π_2 be as in (4.4). Denote

$$\pi_1^* = K_1\lambda_1^*/\lambda_1.$$

We show that if $\lambda_1 > 2\lambda_2\sqrt{c_2}\|\beta^o\|/(c_1+\lambda_2)$, then

$$\mathrm{P}\big(2\sqrt{c^*}\,\tilde{\zeta}(y;m,\mathcal{O},\lambda_2) > \lambda_1\big) \le \pi_1^* + \pi_3. \tag{7.9}$$

Therefore, by Theorem 1, we have

$$\mathrm{P}(\mathrm{sgn}(\hat{\beta}) \neq \mathrm{sgn}(\beta^o) \text{ or } \hat{\beta}(\lambda) \neq \hat{\beta}^o(\lambda_2)) \leq \pi_1 + \pi_1^* + \pi_2 + \pi_3. \tag{7.10}$$

Then Theorem 2 follows from this inequality.

We now prove (7.9). By the definition of $\tilde{P}$,

$$\|(\tilde{P}_B - \tilde{P}_{\mathcal{O}})\tilde{y}\|_2^2 = y'\{Z_B(\Sigma_B + \lambda_2 I_B)^{-1}Z_B' - Z_{\mathcal{O}}(\Sigma_{\mathcal{O}} + \lambda_2 I_{\mathcal{O}})^{-1}Z_{\mathcal{O}}'\}y, \tag{7.11}$$

where $Z_B = n^{-1/2}X_B$. Let $P_B(\lambda_2) = Z_B(\Sigma_B + \lambda_2 I_B)^{-1}Z_B'$ and write $P_B = P_B(0)$. We have

$$\|(\tilde{P}_B - \tilde{P}_{\mathcal{O}})\tilde{y}\|_2^2 = \|(P_B - P_{\mathcal{O}})y\|_2^2 + y'(P_B(\lambda_2) - P_B)y - y'(P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})y. \tag{7.12}$$

Let $T_{B1} = \|(P_B - P_{\mathcal{O}})y\|_2^2$ and $T_{B2} = y'(P_B(\lambda_2) - P_B)y - y'(P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})y$. Let $\eta = \lambda_1/(2\sqrt{c^*})$. Note that $(P_B - P_{\mathcal{O}})y = (P_B - P_{\mathcal{O}})\varepsilon$, since $y = X_{\mathcal{O}}\beta^o + \varepsilon$ and $\mathcal{O} \subseteq B$. Therefore, $T_{B1} = \|(P_B - P_{\mathcal{O}})\varepsilon\|^2$.

Consider T_{B2}. Since $y = X_B\beta_B^o + \varepsilon$, some algebra shows that

$$y'(P_B(\lambda_2) - P_B)y = n\beta_B^{o\prime}Z_B'(P_B(\lambda_2) - P_B)Z_B\beta_B^o + 2\sqrt{n}\beta_B^{o\prime}Z_B'(P_B(\lambda_2) - P_B)\varepsilon + \varepsilon'(P_B(\lambda_2) - P_B)\varepsilon,$$

and $n\beta_B^{o\prime}Z_B'(P_B(\lambda_2) - P_B)Z_B\beta_B^o = -n\lambda_2\|\beta_B^o\|^2 + n\lambda_2^2\beta_B^{o\prime}\Sigma_B^{-1}(\lambda_2)\beta_B^o$. These two equations and the identity $\|\beta_B^o\|^2 - \|\beta_{\mathcal{O}}\|^2 = 0$ imply that $T_{B2} = S_{B1} + S_2 + S_{B3} + S_{B4}$, where

$$\begin{aligned}
S_{B1} &= 2\sqrt{n}\{\beta_B^{o\prime}Z_B'(P_B(\lambda_2) - P_B) - \beta_{\mathcal{O}}^{o\prime}Z_{\mathcal{O}}'(P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})\}\varepsilon, \\
S_2 &= \varepsilon'\{P_{\mathcal{O}} - P_{\mathcal{O}}(\lambda_2)\}\varepsilon, \\
S_{B3} &= \varepsilon'\{P_B(\lambda_2) - P_B\}\varepsilon, \\
S_{B4} &= n\lambda_2^2\{\beta_B^{o\prime}\Sigma_B^{-1}(\lambda_2)\beta_B^o - \beta_{\mathcal{O}}^{o\prime}\Sigma_{\mathcal{O}}^{-1}(\lambda_2)\beta_{\mathcal{O}}^o\}.
\end{aligned}$$

Using the singular value decomposition, it can be verified that $S_{B3} \leq 0$. Also, since $\beta_B^o = (\beta_{\mathcal{O}}^{o\prime}, 0_{|B|-d^o}')'$ and by the formula of the block matrix inverse, it can be verified that $S_{B4} \leq 0$. Therefore,

$$T_{B1} + T_{B2} \leq T_{B1} + |S_{B1}| + S_2. \tag{7.13}$$

Note that $S_2 \geq 0$. When $\alpha = 2$, by Lemma 2 and Proposition 3 of Zhang (2010),

$$\mathrm{P}(\max_{B:|B|=m+d^o} T_{B1} > mn\lambda_1^2/(4c^*)) \leq K_1\frac{2\sqrt{c^*}\sqrt{m}\{m\log(p-d^o)+1\}^{1/\alpha}}{\sqrt{m}\sqrt{n}\lambda_1}.$$

When $1 \leq \alpha < 2$, since $P_B - P_{\mathcal{O}}$ is a rank m projection matrix and there are $\binom{p-d^o}{m}$ ways to choose B from $\{1,\ldots,p\}$, by Lemma 1,

$$\begin{aligned}
\mathrm{P}(\max_{B:|B|=m+d^o} T_{B1} > mn\lambda_1^2/(4c^*)) &\leq K_1\frac{\alpha_n 2\sqrt{c^*}\sqrt{m}\log^{1/\alpha}(m\binom{p-d^o}{m})}{\sqrt{m}\sqrt{n}\lambda_1} \\
&= K_1\frac{\alpha_n 2\sqrt{c^*}\log^{1/\alpha}(m\binom{p-d^o}{m})}{\sqrt{n}\lambda_1}, \\
&\leq K_1\frac{\alpha_n 2\sqrt{c^*}\{m\log(p-d^o+1)\}^{1/\alpha}}{\sqrt{n}\lambda_1},
\end{aligned}$$

where K_1 is a constant that only depends on the tail probability of the error distribution in (A2b). Here we used the inequality $\log(\binom{p-d^o}{m}) \le m\log(e(p-d^o)/m)$.

Let $\mu^o = \sqrt{n}Z_{\mathcal{O}}\beta^o_{\mathcal{O}}$. Since $Z_B\beta^o_B = Z_{\mathcal{O}}\beta^o_{\mathcal{O}} = \mu^o/\sqrt{n}$, we have $S_{B1} = 2\mu^{o\prime}(P_B(\lambda_2) - P_B - (P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})\}\varepsilon$. Write $S_{B1} = 2\|a_B\|(a_B/\|a_B\|)'\varepsilon$, where

$$\|a_B\| = \|\{P_B(\lambda_2) - P_B - (P_{\mathcal{O}}(\lambda_2) - P_{\mathcal{O}})\}\mu^o\| \le \frac{2\lambda_2\|\mu^o\|}{c_* + \lambda_2}.$$

Therefore,

$$\begin{aligned}
\mathrm{P}(\max_{B:|B|=m+d^o} S_{B1} > mn\lambda_1^2/(8c^*)) &\le \mathrm{P}\Big(\frac{4\lambda_2\|\mu^o\|}{c_*+\lambda_2}\max_{B:|B|=m+d^o}|(a_B/\|a_B\|)'\varepsilon| > \frac{mn\lambda_1^2}{8c^*}\Big) \\
&\le K_1\alpha_n\frac{32c^*\|\mu^o\|\lambda_2\log^{1/\alpha}(\binom{p-d^o}{m})}{mn\lambda_1^2(c_*+\lambda_2)}, \\
&\le K_1\alpha_n\frac{32c^*\|\mu^o\|\lambda_2 m^{1/\alpha}\{\log(p-d^o+1)\}^{1/\alpha}}{mn\lambda_1^2(c_*+\lambda_2)}.
\end{aligned}$$

By assumption, $\lambda_2\|\mu^o\| \le \lambda_1/2(c_1+\lambda_2) \le \lambda_1/2(c_*+\lambda_2)$, thus

$$\mathrm{P}(\max_{B:|B|=m+d^o} S_{B1} > mn\lambda_1^2/(8c^*)) \le K_1\alpha_n\frac{16c^*m^{1/\alpha}\{\log(p-d^o+1)\}^{1/\alpha}}{mn\lambda_1(c_*+\lambda_2)^2}. \tag{7.14}$$

For S_2, by Lemma 1,

$$\mathrm{P}(S_2 > mn\lambda_1^2/(8c^*)) \le K_1\alpha_n\frac{8c^*\sigma\lambda_2\sqrt{d^o}\log^{1/\alpha}(d^o+1)}{mn(c_*+\lambda_2)}. \tag{7.15}$$

Inequality (7.10) follows from (7.13) to (7.15). □

References

Breheny, P. & Huang, J. (2009). Coordinate descent algorithms for nonconvex penalized regression methods. *Technical Report #403*, Department of Biostatistics, University of Kentukey.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statist. Sci.* **16**, 199-215.

Chen, S. S., Donoho, D. L. & Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**, 3361.

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

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

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

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

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

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

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

Mazumder, R., Friedman, J. & Hastie, T. (2009). *SparseNet*: Coordinate descent with non-convex penalties. *Tech Report.* Department of Statistics, Stanford University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B.* **58**, 267-88.

Wu, T. & Lange, K. (2007). Coordinate descent procedures for lasso penalized regression. *Ann. Appl. Statist.* **2**, 224-244.

Yuan, M. & Lin, Y. (2007) On the nonnegative garrote estimator. *J. R. Statist. Soc. B.* **69**, 143-161.

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

Zhang, C.-H. & 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 - 2563.

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

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

Table 1: Genes identified using the Enet and the proposed Mnet approach: gene names and nonzero estimates.

Gene name	Enet	Mnet	Gene name	Enet	Mnet
1367731_at	0.0156	0.0402	1378765_at	-0.0013	-0.0020
1368701_at		-0.0301	1379023_at	0.0106	
1368887_at	-0.0012		1379285_at	0.0004	
1368958_at	-0.0443	-0.0490	1379818_at	0.0078	0.0209
1369152_at		0.0250	1380050_at	0.0100	0.0366
1369484_at		0.0251	1380951_at		0.0235
1369718_at	-0.0016		1381508_at	0.0065	
1370434_a_at	-0.0083	-0.0223	1382193_at		0.0208
1370694_at	0.0002		1382365_at		-0.0093
1371052_at	0.0157	0.0332	1385925_at	-0.0004	
1372975_at		-0.0209	1391262_at		0.0209
1373005_at	-0.0116		1392613_at		-0.0022
1375426_a_at		0.0467	1393555_at	-0.0085	
1376129_at	0.0164	0.0536	1394430_at	0.0085	
1376568_at		0.0321	1394459_at	0.0041	
1377651_at	0.0125	0.0404	1394689_at	0.0069	
1387060_at	0.0173	0.0243	1394709_at		0.0003
1387366_at	0.0021		1394820_at	0.0252	
1387902_a_at	0.0059		1395172_at		0.0041
1389795_at	0.0043	0.0107	1396743_at	0.0016	
1390238_at		-0.0270	1397361_x_at	-0.0255	
1390643_at		-0.0281	1398594_at	0.0125	
1378003_at	0.0014				