Minimal class of models for high-dimensional data

Daniel Nevo and Ya'acov Ritov

Abstract

Model selection consistency in the high-dimensional regression setting can be achieved only if strong assumptions are fulfilled. We therefore suggest to pursue a different goal, which we call class of minimal models. The class of minimal models includes models that are similar in their prediction accuracy but not necessarily in their elements. We suggest a random search algorithm to reveal candidate models. The algorithm implements simulated annealing while using a score for each predictor that we suggest to derive using a combination of the Lasso and the Elastic Net. The utility of using a class of minimal models is demonstrated in the analysis of two datasets.

1 Introduction

High dimensional statistical problems have been arising as a result of the vast amount of data gathered today. More specific problem is that estimation of the usual linear regression coefficients vector cannot be performed in the case the number of predictors exceed the number of the observations. In this setting, the usual estimator is irrelevant since the Gram matrix is non invertible. Therefore, a sparsity assumption is often added. For example, the number of regression coefficients that are not equal to zero is assumed to be small. If it was known in advance which predictors has non zero coefficients, the classical linear regression estimator could have been used. Unfortunately, it is not known. Even worse, the natural relevant discrete optimization problem is usually not computationally feasible.

The Lasso estimator, Tibshirani (1996), which solves the problem of minimizing prediction error together with a \( \ell_1 \)-norm penalty, is possibly the most popular method to address this problem, since it results in a sparse estimator. Various algorithms are available to compute this estimator [e.g., Friedman et al. (2010)]. The theoretical properties of the Lasso have been researched throughly in the last decade. For the high-dimension problem, predication rates were established in various manners, Greenshtein & Ritov (2004), Bunea et al. (2006), Bickel et al. (2009), Bunea et al. (2007), Meinshausen & Yu (2009). The capability of the Lasso to choose the correct model depends on nature of the true coefficient vector and the matrix of the predictors, or more precisely, on its Gram matrix, Meinshausen & Bühlmann (2006), Zhao & Yu (2006), Zhang & Huang
However, the underlying assumptions are typically rather restrictive, and cannot be checked in practice.

Many modifications of the Lasso were suggested to overcome its initial disadvantages. For example, the Adaptive Lasso, [Zou (2006)], is a two stage procedure with a second step weighted Lasso, that is, some predictors get less penalty than others: When a grouped structure of the predictors is assumed, the Group Lasso, [Yuan & Lin (2006)], is often used; The Elastic Net estimator, [Zou & Hastie (2005)], is intended to deal with correlated predictors. It is obtained by adding a penalty on the $\ell_2$-norm of the coefficients vector together with the $\ell_1$ Lasso penalty. [Zou & Hastie (2005)] also claim that comparing to the Lasso, prediction using Elastic Net is better, empirically.

In the high-dimensional setting, the task of finding the true model might be too ambitious, if meaningful at all. Only in certain situations, which could not be identified in practice, model selection consistency is guaranteed. Even for the classic setup with more observations than predictors there is no model selection consistent estimator without making further assumptions. This leads us to present a different objective. Instead of looking for the one true model, we aim to present a number of possible models a researcher should look at. Our goal, therefore, is to find models which are almost equivalent in terms of prediction, and hence each one of them should be regarded as a potential good model. Since data is not generated by computer following one's model, there is a benefit in finding several models that their performance is similar, if they are exist. In short, we suggest to find the best models for each reasonable model size. Then, by looking at these models one may reach interesting conclusions regrading the underlying problem. Some of these, as we do below, can be concluded using statistical reasoning, but most of these should be reasoned by a subject matter expert.

Methods that use multiple models for estimation of a parameter of interest are well established in the general statistical framework. This approach was suggested both for the frequentist, [Hjort & Claeskens (2003)], and for the Bayesian, [Hoeting et al. (1999)] statistician. The well known “Bagging”, [Breiman (1996)], is also a technique to combine results from various models. Averaging across estimates obtained by multiple models is usually carried out to account for the uncertainty in the model selection process.

In order to find these models, we implement a search algorithm that uses simulated annealing, [Kirkpatrick et al. (1983)], to find the best models for a given model size. The algorithm is provided with a “score” for each predictor. We suggest to get these scores using a multi-step procedure that implement both the Lasso and the Elastic Net (and then the Lasso again). Multi-step procedures in the high-dimensional setting have drawn some attention and were demonstrated to be better than using the Lasso solely, [Zou (2006)], [Bickel et al. (2010)].

The rest of the paper is organized as follows. Section 2 presents the concept of minimal class of models and the notations. Section 3 provides examples in which the Lasso fails to find the best model. Section 4 described a search algorithm to uncover the relevant models. Section 5 includes a simulation study.
Section 6 wraps up this paper, while technical proofs and supplementary data are included in the appendix.

2 Description of the problem

We start with notations. First, denote \( |v|_q := (\sum |v_q|^q)^{1/q} \) for the \( \ell_q \) (pseudo)norm of any vector \( v \), \( |v|_0 = \lim_{q \to 0} |v|_q \), i.e., the cardinality of \( v \). The data consist of \( X_{n \times p} = (x_1 \ x_2 \ ... \ x_p) \), fixed predictors (also called explanatory variables) matrix, and \( Y_{n \times 1} \), a response vector. WLOG, \( X \) is centered and scaled (\( 1^T x_j = 0 \) and \( |x_j|^2 = 1 \) for all \( j \)) and \( Y \) is centered as well. We are mainly interested in the case \( p > n \). The underlying model is \( Y = X\beta + \epsilon \) where \( \epsilon \) is a random error, \( E(\epsilon) = 0 \), \( V(\epsilon) = \sigma^2 I \). \( \beta \) is the unknown parameter and its true value is denoted by \( \beta^0 \).

Denote \( S \subseteq \{1, ..., p\} \) for a set of indices of \( X \). We call \( S \) a model. We use \( s = |S| \) to denote the cardinality of the set \( S \). Denote also \( S_0 := \{ j : \beta^0_j \neq 0 \} \) and \( s_0 = |S_0| \) for the true model, and its size, respectively. For any model \( S \), we define \( X_S \) to be the submatrix of \( X \) which includes only the columns specified by \( S \). Let \( \hat{\beta}^L_{LS} \) to be the usual least square (LS) estimator of model \( S \), that is, \( \hat{\beta}^L_{LS} := (X^T_S X_S)^{-1} X^T_S Y \), provided \( X^T_S X_S \) is non singular.

Now, the straightforward approach to estimate \( S_0 \) given a model size \( \kappa \) is to consider the following optimization problem:

\[
\min_{\beta} \frac{1}{n} ||Y - X\beta||_2^2, \quad \text{s.t. } ||\beta||_0 = \kappa.
\]  

Unfortunately, typically, this is computationally unfeasible. Therefore, other methods were developed and are commonly used. These methods produce sparse estimators and can be implemented relatively fast. We first present here the Lasso, \textit{Tibshirani (1996)} defined as

\[
\hat{\beta}^L = \arg\min_{\beta} \left( \frac{1}{n} ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \right)
\]  

where \( \lambda > 0 \) is a tuning constant. For some applications, a different amount of regularization is applied for each predictor. This is done using the weighted Lasso, defined by

\[
\hat{\beta}^L_w = \arg\min_{\beta} \left( \frac{1}{n} ||Y - X\beta||_2^2 + \lambda ||w \cdot \beta||_1 \right)
\]  

where \( w \) is a vector of \( p \) weights, \( w_j \geq 0 \) for all \( j \), and \( a \cdot b \) is the Hadamard (Schur, entrywise) product of two vectors \( a \) and \( b \). The Adaptive Lasso, \textit{Zou [3]}
is one example of using a weighted Lasso type estimator. The Elastic Net estimator:

\[
\hat{\beta}^{EN} = \arg\min_{\beta} \left( \frac{1}{n}||Y - X\beta||_2^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2^2 \right). 
\]

(4)

This estimator is often described as a compromise between the Lasso and the well known Ridge regression, Hoerl & Kennard (1970), since it could be rewritten as

\[
\hat{\beta}^{EN} = \arg\min_{\beta} \left( \frac{1}{n}||Y - X\beta||_2^2 + \lambda(\alpha||\beta||_1 + (1 - \alpha)||\beta||_2^2) \right). 
\]

(5)

Let \( \hat{\beta}_n \) and \( \hat{S}_n \) be a sequence of estimators for \( \beta \) and a sequence of the appropriate models, respectively. Model selection consistency is commonly defined as

\[
\lim_{n \to \infty} P(\hat{S}_n = S_0) = 1. 
\]

(6)

If \( p < n \), then criteria based methods e.g., BIC, Schwarz (1978) can be model selection consistent if \( p \) is fixed or if suitable conditions are fulfilled, c.f., Wang, Li & Leng (2009) and references therein. However, these methods are rarely computationally feasible for large \( p \). For \( p > n \), it turns out that strong conditions are needed to achieve (6) for popular regularization based estimators such as the Lasso: Zhao & Yu (2006) and Meinshausen & Bühlmann (2006), the Adaptive Lasso: Huang, Ma & Zhang (2008), the Elastic Net: Jia & Yu (2010) but also for Orthogonal Matching Pursuit (OMP), which is essentially forward selection: Tropp (2004) and Zhang (2009). Furthermore, in Section 3 we provide new examples in which the Lasso fails to find the best model. In particular, we examine the shadowing effect of the Lasso (Section 3.2). There are situation in which an inferior model is chosen by the Lasso and a better model, with equal number of predictors, is hidden. We also give an example where any non-exhaustive algorithm is bound to fail (Section 3.3).

In light of these past results, and the examples given in Section 3, finding several models that are similar in their predication capability could be a more appropriate target. Our goal, therefore, is to find a not too large set of models that predict \( Y \) well enough. Since we are looking for sparse solutions, only the scenario \( \kappa \leq n \) is of interest. For any \( S \) such that \( s \leq n \), it is known that \( \hat{\beta}_S^{LS} \) minimize the sample predication error for a given model \( S \). Thus, this estimator is used for each of the considered models. Finally, \( \mathcal{G} = \mathcal{G}(\kappa, \eta) \) is called a minimal class of models of size \( \kappa \) and efficiency \( \eta \) if

\[
\mathcal{G} = \left\{ S : |S| = \kappa \& \frac{1}{n}||Y - X_S\hat{\beta}_S^{LS}||_2^2 \leq \min_{|S'| = \kappa} \left\{ \frac{1}{n}||Y - X_{S'}\hat{\beta}_S^{LS}||_2^2 \right\} + \eta \right\}.
\]

One could control how similar the models in \( \mathcal{G} \) are to each other in terms of predication, using the tuning parameter \( \eta \). A reasonable way to do so is to choose \( \eta = c\sigma^2 \) with some \( c > 0 \) if \( \sigma^2 \) is known, or to replace \( \sigma^2 \) with an estimate, e.g., \( \hat{\sigma}^2 = \min_{S' \in \mathcal{G}} \left\{ \frac{1}{n}||Y - X_{S'}\hat{\beta}_S^{LS}||_2^2 \right\} \) if \( \sigma^2 \) is unknown.
Note that $G$ depends on $\kappa$, the desired model size. However, in practice one may wish to find $G$ for a few values of $\kappa$, e.g., $\kappa = 1, \ldots, 10$ and then look at all the obtained models. Another option is to replace the Mean Square Error (MSE) $||Y - X_S\hat{\beta}^LS||_2^2$ in the definition of $G$ with one of the available model selection criteria, e.g., AIC [Akaike (1974)] BIC or Lasso. Then, one may consider $\bigcup_{j=1}^K G(j, \eta)$. An alternative to $G$ is to generate the set of models by simply choosing for each $\kappa$ the top $M$ models having the smallest sample MSE.

The next natural question is how should one use a minimal class of models. Examining the models in $G$ may allow us to derive conclusions regarding the importance of different explanatory variables. If, for example, a variable appears in all the models in $G$, we may infer that it is essential for prediction of $Y$, and cannot be replaced. We demonstrate this kind of analysis in Section 5.2.

### 2.1 Relation to other work

A similar point of view on the relevance of a variable was given by Bickel & Cai [2012]. They considered a variable to be important if its relative contribution to the predictive power of a set of variables is high enough. Their next step was to consider only specific type of sets, such that their predication error is not too high, but they also do not contain too many variables.

Rigollet & Tsybakov [2012] investigated a relevant question. They show that linear aggregation of estimators is beneficial in the case of high-dimensional regression when assuming sparsity of the number of used estimators. They also show that the choice of exponential weights corresponds to minimizing a specific, yet relevant, penalized problem.

### 3 Counter examples

The purpose of this section is to show that examples can be constructed in which algorithm like Lasso is not going to find the best models. Relaxing the penalty constant will not reveal the relevant variables. Moreover, any non-exhaustive search algorithm (random or not) is going to fail.

#### 3.1 The Lasso may fail to find the sparser model

We want to argue that the Lasso may find variables which do not reduce the residual variance. The Lasso minimizes the error variance plus penalty on the $\ell_1$ norm, not on the dimension of the model. It may, therefore find not the sparsest possible model. This is known. We want to argue that it actually may add variables which have zero prediction power in the presence of the existing variables.

To make the discussion more transparent, the discussion is checking the population version of the Lasso, where averages are replaced by expectations. However, the parameter $\lambda$ is kept at its sample value (in particular it depends
on \( n \). Let \( \ell = \lambda / n \). Thus we consider the \( y = x^T \beta + \varepsilon = \mu_0 + \varepsilon \), where \( \varepsilon \) is a Gaussian random variable and \( x \) is a random vector of length \( p \). For simplicity \( y \) and the \( x \)'s are assumed to have mean 0 and variance 1. The Lasso is defined to be \( \beta^L := \arg \min_\beta E( y - x^T \beta)^2 + \ell \| \beta \|_1 \) where \( \ell \) is a small number.

Suppose \( y = (x_1 + x_2)/3 + \varepsilon \), and \( x_1, x_2 \) are independent of \( x_3, \ldots, x_p \). The Lasso easily finds the right vector, \( \beta^L = (1/3, 1/3, 0, \ldots, 0)^T \), and \( \| \beta^L \|_1 = 2/3 \).

Suppose now that \( x_1 \) and \( x_2 \) have correlation \( \rho = -8/9 \), and a new variable is introduced \( x_0 = \omega(x_1 + x_2) + \nu \), where \( \omega \in (2/3, 3/4) \), and \( \nu \) is independent of all variables introduced so far and its variance is such that \( \text{Var}(x_0) = 1 \). Let \( \beta' = (\rho, 1/3 - \rho \omega, 1/3 - \rho \omega, 0, \ldots, 0)^T \) (for the new set of predictors, \( x' = (x_0, x_1, \ldots, x_p)^T \)).

Hence

\[
\text{Var}(y - x^T \beta^L) + \ell \| \beta^L \|_1 = \text{Var}(\varepsilon) + 2 \ell / 3
\]

while

\[
\text{Var}(y - x'^T \beta') + \ell \| \beta' \|_1 = \text{Var}(\varepsilon) + \rho^2 \text{Var}(\nu) + \ell (2|1/3 - \rho \omega| + |\rho|)
\]

which is minimized by \( \rho_0 = \ell (2 \omega - 1)/2 \text{Var}(\nu) \). Thus, by adding a new potential variable, the residual variance of the estimator is increased by \( \rho_0 \text{Var}(\nu) \).

### 3.2 The shadowing effect of the Lasso

The following paradoxical situation will be more paradoxical if we consider the constraint version of the Lasso. We present an example where there are two alternative models, one is significantly better than the other (in terms of least squares), yet the inferior model is shadowing its superior, and it can not be observed, even if the better model falls within the constraint. Suppose we have two alternative sparse models, \( S_{\beta'} \) and \( S_{\beta} \), and all variables not in \( S_{\beta'} \cup S_{\beta} \) are just white noise. Suppose \( S_{\beta'} \) have slightly better explanatory power than \( S_{\beta} \), but \( \| \beta' \|_2 = 2 \| \beta^* \|_2 \). Suppose that the difference between them is large enough so that the test between them has negligible sum of errors. When we look for \( \beta \) that minimizes of the sum of squares subject to \( \| \beta \|_1 \leq c \) for \( c \) such that \( \| \beta^* \|_1 \leq c < \| \beta \|_1 \), we expect the minimizer to be approximately \( \beta^* \). However, when the constraint is increased beyond \( \| \beta' \|_1 \), we expect the Lasso solution to be approximately \( \beta_1 \). We describe now a situation in which this is not the case.

Suppose \( y = \xi + \varepsilon \), \( \xi \) and \( \varepsilon \) are independent real random variables, and

\[
x_1 = \frac{1}{2} \xi + \nu
\]
\[
x_2 = \frac{1}{2} \xi - \nu
\]
\[
x_3 = \xi + \zeta + \psi
\]
\[
x_4 = \xi - \zeta + \psi,
\]

where \( y, x_1, \ldots, x_p \) have variance 1, and \( \xi, \psi, \nu, \) and \( \zeta \) are independent, \( \text{Var}(\psi) = n^{-1/2} \). By convexity and symmetry we should consider only vectors \( \beta \) with equal
coefficients to $x_1$ and $x_2$, and equal coefficients to $x_3$ and $x_4$, and 0 otherwise: 

$$\beta = (\beta_1, \beta_1, \beta_2, \beta_2, 0, \ldots, 0)^T.$$ 

Thus, the Lasso solution is the minimizer of 

$$(1 - \beta_1 - 2\beta_2)^2 \text{Var}(\xi) + 4\beta_2^2 \text{Var}(\psi) + 2\lambda(\beta_1 + |\beta_2|).$$ 

which for 

$$\lambda \in \left( \frac{2 \text{Var}(\xi) \text{Var}(\psi)}{3 \text{Var}(\xi) + 2 \text{Var}(\psi), \text{Var}(\xi)} \right)$$ 

is $\beta_1 = 0$, and 

$$\beta_2 = \frac{2 \text{Var}(\xi) - \lambda}{4 \text{Var}(\xi) + 4 \text{Var}(\psi)}.$$ 

Suppose now that $\text{Var}(\xi)$ is of order 1, while $\text{Var}(\psi) = \sqrt{\log n/n}$. In that case the standard Lasso solution would be as above. If $\lambda$ would be smaller so that (7) would not be satisfied, then many of the other $p - 4$ coefficients would be different from 0, but not $\beta_1$. In terms of the constraint version, increasing the constraint to permit $(\beta_1, \beta_2, 0, \ldots, 0)^T$, would actually end with a vector with many nonzeros coefficient, but not the first two.

### 3.3 An example in which the Lasso and heuristic search model are failing

We consider now a variation of the previous model. Let, as before $y = \xi + \varepsilon$. Let $\nu_1, \nu_2, \ldots$ be a sequence of i.i.d. variables independent of $\xi$ and $\varepsilon$, the variance of all these variables is $O(1)$. Let $\psi$ be another independent variable, $\text{Var}(\psi) = o_p(\sqrt{\log p/n})$ but $\text{Var}(\psi) \gg \log p/n$. Let:

$$
x_0 = \xi + \psi$$
$$x_1 = \xi + \nu_1 + \psi,$$
$$x_j = \nu_j - \nu_{j-1}, \quad j = 1, \ldots, k - 1$$
$$x_k = -\nu_{k-1} - \psi$$
$$x_j = \nu_j, \quad j = k + 1, \ldots, p.$$

The true model is $\sum_{j=2}^{k+1} x_j = \xi$. In fact, $x_0$ and $x_1$ are the only explanatory variables correlated with $y$, and the coefficient of $x_1$ in the regression of $y$ on $(x_0, x_1)$ is 0. The only variable that, in the population, has any prediction power to add to $x_0$ is actually the all noise $x_k$. However, its correlation with the residual after regressing $y$ on $x_0$ is $\text{Var}(\psi)/\text{Var}(\xi + \psi)$, which is below the empirical correlation of too many of the nuisance explanatory variables $x_{k+1}, \ldots, x_p$, and by the KKT equation of the Lasso it would likely remain outside the Lasso active set for any reasonable $\lambda$.

This example shows that the Lasso prediction error is $O_p(\sqrt{\log p/n})$ while the optimal achievable rate is $O_p(\log p/n)$, as proved by Rigollet & Tsybakov.
However, the optimal estimator suggested by these authors is numerically infeasible.

We conclude that for this model, any greedy search algorithm that will not “accidentally” hits a model with all of $x_1, \ldots, x_k$, is going to fail. No partial subset of these variable is better than $x_0$ alone, and any subset adds to prediction of $x_0$ less than too many of the nuisance variables.

4 A search algorithm

4.1 Simulated annealing algorithm

In this section, we suggest an algorithm to find $G$ for a given $\kappa$ and $\eta$. The problem is that $||Y - XS\hat{\beta}_S^L||_2^2$ is unknown, and since $p$ is large, even for a relatively small $\kappa$, the number of possible models is huge (e.g., for $p = 200, k = 4$ there are almost 65 million possible models). We therefore suggest to focus our attention on smaller set of models, denoted by $M(\kappa)$. $M$ is a large set of models, but not too large so we can calculate MSEs for all the models within $M$ in a reasonable computer running time. Once we have $M$ and the corresponding MSEs, we can identify $G$ within the models in $M$.

The remaining question is how to assemble $M$ for a given $\kappa$. Any greedy algorithm is bound to find models that are all very similar, concentrating in a specific subspace of the parameter space. Our purpose is to find models that are similar in their predictive power, but heterogeneous in their structure.

Our approach therefore is to implement a search algorithm which travels between potentially attractive models. More specifically, we use a simulated annealing algorithm. The simulated annealing algorithm was suggested for function optimization by Kirkpatrick et al. (1983). The maximizer of a function $f(\theta)$ is of interest. Let $T = (t_1, t_2, \ldots, t_\ell)$ be some decreasing ordered set of “temperatures”. For every temperature level $t \in T$, iterative steps are carried out, before moving for the next, lower, temperature level. These steps involve random suggested move from the current $\theta$ to another $\theta' \neq \theta$ and accepting the move with probability that depends on the ratio $\exp\left[\frac{(f(\theta') - f(\theta))}{t}\right]$ and on the random mechanism of the suggestion. One possible choice for these iterative steps, that is also considered here, are Metropolis-Hastings type steps, Metropolis et al. (1953); Hastings (1970). Then, after some predetermined number of iterations $N_t$, we move to the next $t' < t$ in $T$, taking the last state in temperature $t$ as the initial state for $t'$. The motivation for using this algorithm is that for high enough “temperatures”, moves that do not improve the target function are possible, so the algorithm does not get stuck in a small area of the parameter space. However, as we lower the temperatures, the algorithm moves to new points almost only if the target function is improved. The name of the algorithm and its motivation comes from annealing in metallurgy (or glass processing), where a strained piece of metal is heated, so that a diffusion of its atoms is possible, and then it cools off so the atoms can settle down in low energy position. See Brooks & Morgan (1995) for a general review of the
algorithm in the context of statistical problems.  

In our case, the parameter of interest is $\beta$, or more precisely, the model $S$. The objective function, that we wish to maximize, is

$$f(S) = -\frac{1}{n}||Y - X_S \hat{\beta}^LS||^2.$$  

We now describe the proposed algorithm in more detail. We use stochastic annealing with Metropolis-Hastings steps as a search mechanism for good models. That is, we are not looking for the settling point of the algorithm, but we follow its path, hope that much of it will be in neighborhood of good models, and find the best models along the path.

We say the algorithm is in step $(t,i)$ if the current temperature is $t \in T$ and the current iteration for this temperature is $i \in \{1,\ldots,N_t\}$. For simplicity, we describe here the algorithm for $N_t = N$ for all $t$. Let $S^i_t$ and $\hat{\beta}^i_t$ be the model and the corresponding LS estimator in the beginning of the state $(t,i)$, respectively. An iteration includes a suggested model $S^{i+}_t$, the LS estimator for this model, $\hat{\beta}^{i+}_t$, and a decision whether to move to $S^{i+}_t$ and $\hat{\beta}^{i+}_t$ or to stay at $S^i_t$ and $\hat{\beta}^i_t$.

We now need to define how does $S^{i+}_t$ is suggested and what is the probability of accepting this move. For each $S^i_t$, we suggest $S^{i+}_t$ by a minor change, i.e., we take one variable out and we add another in, and then obtain $\hat{\beta}^{i+}_t$ by standard linear regression. Assume that for every variable $j \in (1,\ldots,p)$ we have a score $\gamma_j$, such that higher $\gamma_j$ reflects that the variable $j$ should be in the model, comparing with other possible variables. WLOG, assume $0 \leq \gamma_j \leq 1$ for all $j$. We choose a variable $r^* \in S^i_t$ and take it out with the probability function

$$p^\text{out}_{t,r} = \frac{\gamma_r - 1}{\sum_{u \in S^i_t} \gamma_u - 1}, \quad \forall r \in S^i_t. \quad (8)$$

Then, we choose a variable $\ell^* \notin S^i_t$ and add it to the model with the probability function

$$p^\text{in}_{t,\ell} = \frac{\gamma_\ell}{\sum_{u \notin S^i_t} \gamma_u}, \quad \forall \ell \notin S^i_t. \quad (9)$$

Thus

$$S^{i+}_t = \{S^i_t \setminus \{r^*\} \cup \{\ell^*\}$$

and we may calculate the LS solution $\hat{\beta}^{i+}_t$ for the model $S^{i+}_t$. The first part of our iteration is over. A potential candidate was chosen. The second part is the decision whether to move to the new point or to stay at the current point. Following the scheme of simulated annealing algorithm with Metropolis-Hastings type moves we calculate

$$q = \exp\left(\frac{1}{n\ell} \left(||Y - X_S \hat{\beta}^{i+}_t||^2_2 - ||Y - X_{S^i_t} \hat{\beta}^{i+}_t||^2_2\right)\right) \frac{p(S^i_t \rightarrow S^{i+}_t)}{p(S^{i+}_t \rightarrow S^i_t)}$$

9
where
\[ p(S_i^t \rightarrow S_i^{t+1}) \equiv p_{i,r_t}^{out} \cdot p_{i,t}^{in}, \]
\[ p(S_i^{t+1} \rightarrow S_i^t) \equiv p_{i+t}^{out} \cdot p_{i+r}^{in}. \] (10)

We are now ready to the next iteration \( i + 1 \) by setting
\[
(S_i^{t+1}, \hat{\beta}_{i}^{t+1}) = \begin{cases} 
(S_i^t, \hat{\beta}_{i}^t) & \text{w.p. } \min(1, q) \\
(S_i^t, \hat{\beta}_{i}^t) & \text{w.p. } \max(0, 1 - q) 
\end{cases}
\]

Along the run of the algorithm, the suggested models and their corresponding MSEs are kept. Then, these models are used to form \( \mathcal{M}(\kappa) \) and \( \mathcal{G} \) can be identified given \( \eta \).

We now point out several issues that should be considered when using the algorithm. First, the algorithm described above for one single value of \( \kappa \). In practice, one may run the algorithm separately for different values of \( \kappa \). Another consideration is the tuning parameters of the algorithm that are provided by the user: The temperatures \( T \); the number of iterations \( N \); the starting point \( S_1^t \); and the vector \( \gamma = (\gamma_1, ..., \gamma_p) \). Our empirical experience is that the first three can be managed without too many concerns. See Section 4. Regarding the vector \( \gamma \), a wise choice of this vector should improve the chance of the algorithm to move in desired directions. We deal with this question in Section 4.2. However, in what follows we show that under suitable conditions the algorithm will work well even for a general choice of \( \gamma \).

Define \( S_0, s_0 \) and \( \beta_0 \) as before and let \( \mu = X\beta_0 \). That is, \( Y = \mu + \epsilon \). Denote \( P_S \) for the projection matrix onto the subspace spanned by the submatrix \( X_S \).

We first introduce few simple and common assumptions:
(A1) \( ||\mu||_2^2 = O(n) \)
(A2) \( s_0 \) is small, i.e., \( s_0 = O(1) \).
(A3) \( p = n^a, a > 1 \)
(A4) \( \epsilon \sim N_n(0, \sigma^2 I) \)

Denote \( A_\gamma \) for the set of positive entries in \( \gamma \). That is, \( A_\gamma \subseteq \{1, 2, ..., p\} \) is a (potentially) smaller group of predictors than all the \( p \) variables. Denote also \( h_\gamma = |A_\gamma| \) for the size of \( A_\gamma \), and \( \gamma_{min} := \min_{i \in A_\gamma} \gamma_i \) for the lowest positive entry in \( \gamma \).

Informally, the algorithm is expected to preform reasonably well if:
1. The true model is relatively small (e.g., with 10 active variables).
2. A variable in the true model is adding to the prediction of a set of variables if a very few (e.g., 2) other variables are in the set.

Our next assumption is more restrictive. Let \( \hat{S} \) be an interesting model with size \( s_0 \). That is, a small model with low MSE. The models we are looking for
are of this nature. We facilitate the idea of \( \tilde{S} \) being an interesting model by assuming that \( X \beta_{\tilde{S}} \) is close to \( \mu \) (in the asymptotic sense). More specifically, we virtually assume that for every model with \( s = s_0 \), which is not \( \tilde{S} \), if we take out a predictor that is not part of \( \tilde{S} \), and replace it with a predictor from \( \tilde{S} \), the subspace spanned by the new model is not much further from \( \mu \), comparing with the subspace spanned by the original model. More formerly, denote \( P_S \) for the projection matrix onto the subspace spanned by the columns of the submatrix \( X_S \).

\[ (B1) \text{ There exist } t_0 > 0 \text{ and a constant } c > 0, \text{ such that for all } S, |S| = s_0 - 1 \text{ and for all } j \in \tilde{S} \cap S^c, j' \in S^c \cap S^c, \text{ and for a large enough } n \]

\[ \frac{1}{n} \left[ \|P_{S^c} \mu\|^2_2 - \|P_{S^c} \mu\|^2_2 \right] > 4t_0 \log c, \]

where \( S^c_r \equiv S \cup \{r\} \).

The following theorem gives conditions under which the simulated annealing algorithm is passing through the model \( \tilde{S} \). More accurately, the theorem states that there is always strictly positive probability to pass through \( \tilde{S} \) in the next few moves. This result should apply for all models that Assumption (B1) holds for. Note however, that we do not claim that the algorithm finds all the models in a minimal class. Proving such a result would probably require complicated assumptions on models with larger size than \( s_0 \), and their relation to \( \tilde{S} \) and other interesting models.

Let \( P^m_t (S'|S) \) be the probability of passing through model \( S' \) in the next \( m \) iterations of the algorithm, given the current temperature is \( t \), and the current state of the algorithm is the model \( S \).

Theorem 4.1 Consider the simulated annealing algorithm with \( \kappa = s_0 \) and with a \( \gamma \) vector such that \( \gamma_{\min} \geq c_\gamma \). Let Assumptions (A1)-(A4) hold and let Assumption (B1) hold for some temperature \( t_0 \) and with \( c = c_\gamma \). If \( \tilde{S} \subseteq A_\gamma \) then for all \( S \subseteq A_\gamma \) with \( s = s_0 \), for all \( m \geq s_0 - |\tilde{S} \cap S| \) and for large enough \( n \),

\[ P^m_{t_0} (\tilde{S}|S) > \left[ \frac{c^2_{\gamma}}{s_0(h_{\gamma} - s_0)} \right]^{s_0}. \] (11)

A proof is given in the appendix. Theorem 4.1 states that for any choice of the vector \( \gamma \), such that the entries in \( \gamma \) are positive for all true model predictors, the probability that the algorithm would visit a specific interesting model in the next \( m \) moves is always positive, provided the temperature is high enough, and provided it is possible to move from the current model to this model in \( m \) moves. While our intention is to use the algorithm as a search algorithm for several models, Theorem 4.1 gives motivation to also use the algorithm to solve the optimization problem (1).

For the classical model selection setting with \( p < n \), a similar method was suggested by Brooks, Friel & King (2003). Their motivation is as follows. When
searching for the most appropriate model, likelihood based criteria are often used. However, maximizing the likelihood to get parameters estimates for each model becomes infeasible as the number of possible models increases. They therefore suggest to simplify the process by maximizing simultaneously over the parameter space and the model space. They suggest a simulated annealing type algorithm to implement this optimization. Their suggested algorithm is essentially an automatic model selection procedure.

4.2 Choosing \( \gamma \)

The simulated annealing algorithm described above is provided with the vector \( \gamma \). The values \( \gamma_1, \ldots, \gamma_p \) should represent the knowledge regarding the importance of the predictors, although we do not assume here that any prior knowledge is available. These values are used for the suggestion of the new point for the algorithm. As it can be seen in equations (8)-(9), predictors with high \( \gamma \) values has larger probability to enter the model if they are not part of the current model, and lower probability to be suggested for replacement if they are already part of it. Since \( p \) is large, we may also benefit if the suggested vector includes many zeros, for predictors that are clearly not affective for prediction of \( Y \).

The simplest way to choose \( \gamma \) is to take the absolute values of the univariate correlation of each predictor with \( Y \), and then we may threshold it and keep only large enough values. However, this choice does not take into account the presence of other predictors when judging whether a predictor is important for prediction of \( Y \).

Another possibility is to obtain an initial estimator for \( \beta \) using the Lasso, and take \( \hat{\beta}^L \) with a relatively large model \( s \approx n \) and then to set \( \gamma_j = \frac{|\beta_j|}{||\hat{\beta}^L||_1} \), since predictors with large coefficient value may be more important.

However, as discussed in Sections 2 and 3, the Lasso might miss some potentially good predictors as a result of the covariance matrix structure. It is well known that the Elastic Net may add these predictors to the solution, although it might also add unnecessary predictors. Moreover, it is not clear how to choose \( \gamma_j \) using solely the Elastic Net. While the Lasso and the Elastic Net estimators are not model selection consistent in many situations, and hence could not be used to our purposes, by combining both methods together we may get a reservoir of hopefully interesting predictors.

Zou & Hastie (2005) provided motivation and results that justify the common knowledge that the Elastic Net is better to use with correlated predictors. Since we intend to exploit this property of the Elastic Net this paper includes some theoretical justification. We present a more general result later on this section, but for now, the following proposition demonstrates why the Elastic Net tends to include correlated predictors in its model.

Proposition 4.2 Define \( X \) and \( Y \) as before, and define \( \hat{\beta}^{EN} \) by (4). Denote \( \rho = x_1^T x_2 \). Assume \( |\hat{\beta}_1^{EN}| \geq c_\beta \) for some \( c_\beta > 0 \). If \( |\rho| > 1 - \lambda_2^2 c_\beta^2 / ||Y||_2 \) then \( |\hat{\beta}_2^{EN}| > 0 \).
A proof is given in the appendix. Proposition 4.2 gives motivation for the $\hat{\beta}^{EN}$ tendency to have larger model than $\hat{\beta}^{L}$. It quantifies how much correlated two predictors need to be so the Elastic Net would either include both predictors or none of them.

So we use all the predictors that are part of the Lasso or the Elastic Net models. The question remains how to give “scores” to the different predictors. Let $S_L$ and $S_{EN}$ be the models obtained by $\hat{\beta}^{L}$ and $\hat{\beta}^{EN}$, respectively. Define $S_+$ for the group of predictors that were part of the Elastic Net model but not part of the Lasso model and $S_{out}$ for the predictors that were not included in any of them. Note that $S_L \cap S_+ = S_L \cap S_{out} = S_+ \cap S_{out} = \emptyset$ and $S_L \cup S_+ \cup S_{out}$ is $\{1, \ldots, p\}$. Define

$$\hat{\beta}^{L+}_j(\delta) = \arg \min_{\beta} \left( ||Y - X\beta||^2 + \lambda \sum_{j=1}^{p} \delta^{I\{j \in S_+\}} |\beta_j| \right), \quad \delta \in (0, 1),$$

and let $S^L_+(\delta)$ be the appropriate model. That is, a reduced penalty is given for predictors that $\hat{\beta}^{L}$ might have missed in the first run. Thus, these predictors are encouraged to enter the model, and since these predictors may take the place of others, predictors in $S_L$ that their explanation power is not high enough are pushed out of the model. Note that $\hat{\beta}^{L+}_L(\delta)$ is a special case of $\hat{\beta}^{L+}_j$ as defined in [3] with $w_j = \delta^{I\{j \in S_+\}}$.

We demonstrate this reduced penalty idea using a toy example. A data set with $n = 30$ and $p = 50$ is simulated. The true value of $\beta$ is taken to be $\beta^0 = (0.5 0.5 1 1 0 0 ... 0)^T$ and $\sigma^2$ is taken to be one. The predictors are independent normal variables with the exception of 0.8 correlation between $x_1$ and $x_2$. Predictor 1 is included in the Lasso model, however predictor 2 is not. Figure 1 presents the coefficients’ estimates of $x_1$, $x_2$ and $x_3$ when lowering the penalty of $x_2$. Note how $x_2$ enters the model for low enough penalty while $x_1$ leaves the model for low enough penalty (on $x_2$).

Using the results of applying reduced penalties, we suggest to measure the importance of a predictor $j \in S_+$ by the highest $\delta$ such that $j \in S^L_+(\delta)$. On the other hand, the importance of a predictor $z' \in S_L$, can be measured by the highest $\delta$ such that $z' \notin S^L_+(\delta)$ (now, smaller $\delta$ reflects $z'$ is more important). With this in our mind, we continue to derive $\gamma$.

Let $\Delta = (\delta_0 < \delta_1 < \ldots < \delta_h)$ be some grid of $[0, 1]$, with $\delta_0 = 0$ and $\delta_h = 1$. For each $\delta \in \Delta$, we obtain $\hat{\beta}^{L+}_L(\delta)$. Define

$$i^{\gamma}_j = \begin{cases} \arg \max_i \{ i : \hat{\beta}^{L+}_L(\delta_i)_j \neq 0 \} & j \notin S_L \\ \arg \max_i \{ i : \hat{\beta}^{L+}_L(\delta_i)_j = 0 \} & j \in S_L \end{cases}$$

and if the arg max is over an empty set, define $i^{\gamma}_j = 0$. Let $\delta^j := \delta^{i^{\gamma}_j}$. Now, we
suggest to choose $\gamma_j$ as follows:

$$
\gamma_j = \begin{cases}
0 & j \in S\text{out} \\
\frac{\delta^i_j}{2} & j \in S_+ \\
1 - \frac{\delta^i_j}{2} & j \in S_L,
\end{cases}
$$

for all $j \in \{1, ..., p\}$. This choice of $\gamma$ has the following nice properties.

- A predictor $j \notin S_L$ with $i^*_j = 0$ is no longer considered to be part of any model.

- On the other hand, for a predictor $j \in S_L$, if $i^*_j = 0$ than $\gamma_j = 1$ which is the maximal possible value. This means that even when the penalty for other predictors was dramatically reduced, leading to their entrance to the model, $j$ is essential for prediction of $Y$.

- Since the predictors in $S_L$ were picked when equal penalty was assigned to all predictors, they get priority over the predictors in $S_+$.

- However, for two identical predictors (or highly correlated predictors), WLOG $x_1 = x_2$ such that $1 \in S_L$ and $2 \notin S_L$, we get a desirable result. Following Proposition 4.2, we know that $X_2 \in S_+$. Now, for $\delta_{h-1} < 1$ it is clear that $2 \in S^L_+(\delta_{h-1})$ and $1 \notin S^L_+(\delta_{h-1})$. Therefore $i^*_1 = i^*_2 = h - 1$, and hence if $\delta_{h-1}$ is taken to be close to one, then $\gamma_1 \simeq \gamma_2 \simeq 0.5$ as one my wish, since $x_1$ and $x_2$ hold a similar predicative power of $Y$.

Proposition 4.2 deals with the case of two correlated predictors. In practice, the covariance structure may be much more complicated. Therefore the question

Figure 1: Toy example: coefficients’ estimates for predictors $x_1, x_2$ and $x_3$ when lowering the Lasso penalty for $x_2$ only. The rightmost point corresponds to a Lasso procedure with equal penalties for all predictors.
arises: can we say something more general on the Elastic Net in the presence of competing models? Apparently we can. Let $M_1$ and $M_2$ be two models, that is, two sets of predictors, that possibly intersect. Assume that the Elastic Net solution chose all the predictors in $M_1$, what can we say about the predictors in $M_2$? Are there conditions on $X_{M_2}$, $X_{M_1}$ and $Y$ such that all the predictors in $M_2$ were also chosen? If the answer is yes (and it is, as Theorem 4.3 states), it justifies our use of the Elastic Net to reveal more relevant predictors. In our case, the relevant predictors are the building blocks of models in $G$.

In order to reveal this property of the Elastic Net, we analyze $\hat{\beta}_{EN}$, the solution of (4), when assuming all the predictors in $M_1$ have non-zero values. We denote $M(-)$ for $M_1 \cap M_2$, the set of predictors that are not included in $M_1$ or $M_2$ and $\tilde{X} = X_{M(-)}$ for the appropriate submatrix of $X$. We let $\hat{\beta}_{EN}^{M_1}$, $\hat{\beta}_{EN}^{M_2}$ and $\hat{\beta}_{EN}^{M_1 \cap M_2}$ be the coordinates of $\hat{\beta}_{EN}$ that correspond to $M_1$, $M_2$ and $M_1 \cap M_2$, respectively. Then, we show that we can concentrate on $\tilde{Y} = Y - \tilde{X}\hat{\beta}_{EN}$, which is the unexplained reminder of $Y$, after taking into account $\tilde{X}$. Finally, we show that both $M_1$ and $M_2$ are chosen by the Elastic Net if the prediction of $\tilde{Y}$ using $M_1$, namely $X_{M_1}\hat{\beta}_{EN}^{M_1}$, projected onto the subspace spanned by the columns of $M_2$ is correlated enough with $\tilde{Y}$. Formally,

**Theorem 4.3** Define $\hat{\beta}_{EN}$ as before. Let $M_1$ and $M_2$ be two models with the appropriate submatrices $X_{M_1}$ and $X_{M_2}$. Define $\tilde{X}$ and $\tilde{Y}$ as before. Define $\hat{\beta}_{EN}^{M_1}$ and $\hat{\beta}_{EN}^{M_2}$ as before. Denote $P_{M_2}$ for the projection matrix onto the subspace spanned by the columns of $X_{M_2}$. WLOG, assume $|M_2| \leq |M_1|$ and that all the coordinates of $\hat{\beta}_{EN}^{M_1}$ are different than zero. Finally, if

$$\tilde{Y}^T P_{M_2} X_{M_1} \hat{\beta}_{EN}^{M_1} > c_1(\lambda_1, \lambda_2, X_{M_1}, \tilde{Y}, \hat{\beta}_{EN}^{M_1}),$$

(12)

then all the coordinates of $|\hat{\beta}_{EN}^{M_2}|$ are different than zero.

A proof and a discussion on the technical aspects of condition (12) and the constant $c_1$ are given in the appendix. Theorem 4.3 states that under a suitable condition, predictors belong to at least one of two competing models are chosen by the Elastic Net. In our context, when we have a model $M_1$ with a good prediction accuracy, i.e., $X_{M_1}\hat{\beta}_{EN}^{M_1}$ is close to $\tilde{Y}$, then predictors in any another model $M_2$ which has similar prediction, that is $P_{M_2} X_{M_1} \hat{\beta}_{EN}^{M_1}$ is also close to $\tilde{Y}$, would be chosen by the Elastic Net. Hence, these predictors are expected to have a positive value in $\gamma$, and our simulated annealing algorithm would pass through these models, provided the conditions in Theorem 4.1 are met. Therefore, these models are expected to appear in $G$. 

15
5 Numerical Results

5.1 Simulation Study

We consider a setup in which there are few models one would want to reveal. The following model is used

\[ Y = X\beta + \epsilon, \quad \epsilon \sim N(0, I) \]

with \( \beta_j \) equals to \( C \) for \( j = 1, 2, \ldots, 6 \) and zero for \( j > 6 \). \( C \) is a constant chosen to get the desired signal to noise ratio (SNR). All the predictors are i.i.d \( N(0, 1) \) besides \( X(7) \) and \( X(8) \) which are defined by

\[
X(7) = \frac{2}{3}[X(1) + X(2)] + \xi_1, \quad \xi_1 \sim N_n \left( 0, \frac{1}{3}I \right)
\]

\[
X(8) = \frac{2}{3}[X(3) + X(4)] + \xi_2, \quad \xi_2 \sim N_n \left( 0, \frac{1}{3}I \right)
\]

where \( \xi_1 \) and \( \xi_2 \) are independent. In this scenario, there are 4 models we would like to find: (I) \{1,2,3,4,5,6\}; (II) \{5,6,7,8\}; (III) \{3,4,5,6,7\}; and (IV) \{1,2,5,6,8\}.

For each simulated dataset, we do the following:

1. Obtain \( \gamma \) as explained in Section 4.2. The tuning parameter of the Lasso is taken to be the minimizer of the cross-validation MSE. For the Elastic Net, \( \alpha \) in (5) is taken to be 0. These values are used so the pool of the possible predictors would be large.

2. Run the simulated annealing algorithm for \( \kappa = 4, 5, 6 \). The tuning parameters of the algorithm are chosen quite arbitrarily: \( T = 10 \times (0.7^1, 0.7^2, \ldots, 0.7^{20}); \Delta = (0, 0.02, 0.04, \ldots, 0.98, 1); N_t = N = 100 \) for all \( t \in T \).

3. For each model (I)–(IV), we check whether the model is the best model obtained (as measured by MSE) within models with the same size. For example, we check if Model (II) is the best model out of all models with \( \kappa = 4 \). We also check whether the model is one of the top five models within models with its size.

We generated 1000 simulated datasets for different scenarios: For \( n = 100, p = 200, 500, 1000 \) and for SNR = 1, 2, 4, 8, 12, 16. Table 1 displays the proportion of times each model was chosen, either as the best one, or as one of the top five models. The results are as one may have expect. For larger SNR, the models are chosen more frequently, and as the SNR grows, the results for choosing the top 5 models are similar to the best model results. However, models (III) and (IV) are competing, in the sense that they both include 5 predictors. Even for large SNR, each of the models (III)-(IV) is chosen in about 50% of the cases. As recommended in Section 4.1, we should start the algorithm from different points, that is, different models. Figure 2 presents comparison between running the algorithm once and three times, from different points. Note the improved results for models (III) and (IV) when the algorithm is used three times. The results described in this section were quite similar when we formed \( \mathcal{G}^*(\kappa, \eta) \) for each \( \kappa = 4, 5, 6 \) separately, using an arbitrary small value of \( \eta \).
Table 1: Proportion that each model is chosen as best model or as one of top five models for different number of potential predictors (p) and various SNR values

| SNR | Model | p = 200 |     | p = 500 |     | p = 1000 |     |
|-----|-------|---------|-----|---------|-----|----------|-----|
|     |       | Best    | Top 5 | Best    | Top 5 | Best    | Top 5 |
| 1   | (I)   | 0.00    | 0.01 | 0.00    | 0.00 | 0.00    | 0.00 |
|     | (II)  | 0.42    | 0.62 | 0.28    | 0.46 | 0.23    | 0.38 |
|     | (III) | 0.04    | 0.08 | 0.01    | 0.02 | 0.00    | 0.00 |
|     | (IV)  | 0.04    | 0.08 | 0.02    | 0.03 | 0.00    | 0.01 |
| 2   | (I)   | 0.10    | 0.12 | 0.05    | 0.06 | 0.04    | 0.05 |
|     | (II)  | 0.94    | 0.96 | 0.92    | 0.94 | 0.94    | 0.95 |
|     | (III) | 0.27    | 0.34 | 0.18    | 0.24 | 0.15    | 0.17 |
|     | (IV)  | 0.28    | 0.37 | 0.18    | 0.22 | 0.14    | 0.17 |
| 4   | (I)   | 0.38    | 0.38 | 0.20    | 0.20 | 0.11    | 0.11 |
|     | (II)  | 0.96    | 0.96 | 0.96    | 0.96 | 0.96    | 0.95 |
|     | (III) | 0.38    | 0.46 | 0.31    | 0.36 | 0.22    | 0.24 |
|     | (IV)  | 0.39    | 0.46 | 0.28    | 0.31 | 0.24    | 0.26 |
| 8   | (I)   | 0.72    | 0.72 | 0.46    | 0.46 | 0.32    | 0.32 |
|     | (II)  | 0.97    | 0.97 | 0.97    | 0.97 | 0.96    | 0.96 |
|     | (III) | 0.41    | 0.48 | 0.36    | 0.40 | 0.30    | 0.31 |
|     | (IV)  | 0.44    | 0.50 | 0.34    | 0.37 | 0.29    | 0.31 |
| 12  | (I)   | 0.86    | 0.86 | 0.66    | 0.66 | 0.49    | 0.49 |
|     | (II)  | 0.98    | 0.98 | 0.97    | 0.97 | 0.96    | 0.96 |
|     | (III) | 0.49    | 0.55 | 0.41    | 0.44 | 0.32    | 0.34 |
|     | (IV)  | 0.42    | 0.48 | 0.37    | 0.40 | 0.32    | 0.34 |

5.2 Real data sets

We demonstrate the utility of our method on two real datasets. The tuning parameters of the Lasso and the Elastic Net were taken to be the same as in Section 5.1. The tuning parameters of the simulated annealing algorithm were $T = 10 \times (0.7^1, 0.7^2, ..., 0.7^{20})$, $\Delta = (0, 0.01, 0.02, ..., 0.98, 0.99, 1)$, and $N_t = N = 100$ for all $t \in T$.

5.2.1 Riboflavin

We use a high-dimensional data about the production of riboflavin (vitamin B2) in Bacillus subtilis that were recently published, B"uhlmann, Kalisch & Meier (2014). The data include $p = 4088$ predictors which are measures of log expression levels of genes in $n = 71$ observations. The target variable is the (log) riboflavin production rate.

$S_L$ included 40 predictors (and intercept), and $S_{EN}$ included 59 predictors when taking the tuning parameters as described in Section 5.1. In total, we considered 61 different predictors (i.e., genes). Panel (a) of Figure 3 presents the histogram of the positive values in $\gamma$.

We run the algorithm from three random starting points for each model.
Figure 2: Proportion that each model is chosen as one of top five models for different number of potential predictors ($p$) and various SNR values. There is an apparent improvement when running the algorithm from three starting points.

Figure 3: Histograms of the values of $\gamma$ for positive entries only in the two dataset analysis examples.
Table 2: Riboflavin data: Number of unique models for each model size after running the algorithm from 3 different starting points

| Model size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|------------|---|---|---|---|---|---|---|---|---|----|
| Number of models | 5 | 5 | 8 | 6 | 13 | 15 | 15 | 15 | 15 | 15 |

size between 1 and 10. We kept the best 5 models for each size and for every different starting point. We then combined these models to get, after removal of duplicates, a total of 112 models. See Table 2 for the number of unique models as a function of the model size. The following insights are drawn from examining more carefully the obtained models (see Table 5 in the appendix):

- In total, the models include 53 different predictors. Out of these, 35 predictors appear in less than 10% of the models, meaning they are probably less important as predictors of riboflavin production rate.

- Gene number 2564 appears in all models of size larger than 3 and in 5 out of 8 models of size 3. However, this gene is not included in any of the smaller models. This gene is the only one that appears in more than half of the obtained models. We can infer that while this gene does not hold an effect strong enough comparing to other genes in order to stand out, it has a unique relation with the outcome predictor that could not be mimicked using other combination of genes.

- At least one gene from the group \{4002, 4003, 4004, 4006\} is contained in the models of size larger than one, although never more than one of these genes. Genes number 4003 and 4004 appear more frequently than genes number 4002 and 4006. Looking at the correlation matrix of only this group of predictors, we see they are all highly correlated (pairwise correlations > 0.97). Future research could take this finding into account by using, e.g., the Group Lasso of Yuan & Lin (2006).

- Similarly, genes number 1278 and 1279 appear in about half of the models, but always only one of them. They are also strongly correlated (0.984). The same statement holds for genes number 69 and 73 (correlation of 0.945) as well.

- Genes number 792,1131 and maybe a few more, could be also considered as they appear in a variety of different models.

We now move to comparison of our results with models obtained using other methods, as reported in Bühlmann et al. (2014). The multiple sample splitting method to get p-values, Meinshausen et al. (2009), yields only one significant predictor. Indeed, a model that includes only this predictor is part of our considered models. If one constructs his model using the stability selection, Meinshausen & Bühlmann (2010) as a screening process for the predictors, he would get a model of three genes, which correspond to columns number
625, 2565 and 4004 in our $X$ matrix. However, this model is not included in our top models. In fact, the highest sample MSE for a model in our 8 models of size 3 is 0.2047 while the sample MSE of the model suggested using the stability selection is 0.2703, more than 30% difference!

5.2.2 Air pollution

We now demonstrate how the proposed procedure can be used for allegedly simpler, traditional, problem. The air pollution data set, McDonald & Schwing (1973) includes 58 Standard Metropolitan Statistical Areas (SMSAs) of the US (after removal of outliers). The outcome variable is the age adjusted mortality rate (per 100k). There are 15 potential predictors including air pollution, environmental, demographic and socioeconomic predictors. The predictors are described in Table 6 in the appendix.

There is no guarantee that the relationship between the predictors and the outcome variable has linear form. We therefore include commonly used transformations of each variable, namely natural logarithm, square root and power of two transformations. Considering also all possible two way interactions, we have a total of 165 predictors. Then, the obtained $X$ matrix is scaled.

The problem of high-dimensional data composed of transformations and interactions has been dealt with in the literature. For example, by using two steps procedures, e.g., Bickel et al. (2010) or by solving optimization problem, e.g., Bien et al. (2013). Our procedure has a different goal, since we are not looking for the best predictive model, but rather for a meaningful insights about the data.

After the Lasso and Elastic Net step, we are left with 44 predictors with positive $\gamma_j$ (one “natural” predictor, 3 log transformations, 4 square root transformations, 8 power of two transformations and the rest are interactions). Panel (b) of Figure 3 presents the histogram of the positive values in $\gamma$.

For every $\kappa = 1, 2, ..., 10$, we run the algorithm from three starting points, each time keeping the 5 best models. In total, we get 126 unique models. Table 3 summarizes the results for prominent predictors, that is, predictors that appear in at least quarter of the obtained models. This table presents a matrix of the joint frequency of each two predictors. Each cell in the table is the number of models such that both the predictor listed in the row and the predictor listed in the column are part of the model. The diagonal is simply the number of models that a predictor appears in.

Three (transformed) main effects are chosen. The nitric oxide pollution is invaluable for prediction of mortality rate. This predictor (in a log shape) appears in a large majority of the models. Apart from this predictor, the hydrocarbon pollution appears (after a square root transformation) but only in about 30% of the models. Before moving forward with the analysis there is one finding that catches the eye. The two zeros in the matrix mean that interactions involving the percentage of non white are only part of models that do not include the
Table 3: Frequency of each two predictors to be selected together in the 126 obtained models. The diagonal is simply the number of models that a predictor appears in. For example, both $\log(\text{NOx})$ and $\sqrt{\text{nwht}}$ appear in 27 models:

|     | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| (1) $\log(\text{NOx})$ | 97  | 27  | 36  | 30  | 50  | 31  | 33  | 35  |
| (2) $\sqrt{\text{nwht}}$ | 33  | 7   | 8   | 14  | 10  | 0   | 0   |     |
| (3) $\sqrt{\text{HC}}$ | 37  | 12  | 18  | 10  | 16  | 15  |     |     |
| (4) $\text{HC} \times \text{prec}$ | 33  | 26  | 17  | 18  | 8   |     |     |     |
| (5) $\text{jant} \times \text{ovr65}$ | 66  | 30  | 27  | 26  |     |     |     |     |
| (6) $\text{pphs} \times \text{educ}$ | 37  | 14  | 14  |     |     |     |     |     |
| (7) $\text{nwht} \times \text{ofwk}$ | 46  | 1   |     |     |     |     |     |     |
| (8) $\text{nwht} \times \text{mst}$ | 43  |     |     |     |     |     |     |     |

Table 4: Frequency of each two predictors to be selected together in the 119 obtained models, after removal of two problematic interactions. The diagonal is simply the number of models that a predictor is included. For example, in 40 models both $\log(\text{NOx})$ and $\sqrt{\text{nwht}}$ appear. The matrix includes all the predictors that are part of at least quarter of the models:

|     | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| (1) $\text{nwht}$ | 37  | 13  | 36  | 0   | 12  | 17  | 28  | 26  |
| (2) $\log(\text{prec})$ | 31  | 31  | 13  | 9   | 2   | 15  | 19  |     |
| (3) $\log(\text{NOx})$ | 106 | 43  | 28  | 40  | 67  | 62  |     |     |
| (4) $\sqrt{\text{nwht}}$ | 44  | 11  | 13  | 24  | 24  |     |     |     |
| (5) $\text{dens} \times \text{prec}$ | 30  | 14  | 26  | 20  |     |     |     |     |
| (6) $\text{hum} \times \text{prec}$ | 40  | 38  | 28  |     |     |     |     |     |
| (7) $\text{jant} \times \text{ovr65}$ | 68  | 47  |     |     |     |     |     |     |
| (8) $\text{pphs} \times \text{educ}$ |     |     |     |     |     |     |     | 63  |

The obtained frequency matrix is displayed in Table 3. The conclusion regarding the importance of the nitric oxide pollution remains. Nevertheless, hydrocarbon pollution is not relevant anymore. The percentage of non-white population appears untransformed but also after taking its squared root. However, this predictor appears in one form only each time. We conclude that this predictor should be used for prediction of the mortality rate, but the question of transformation remains unsolved.

Turning to the interactions. The interaction between percentage of elderly population and the average temperature in January appears while the appropriate main effects are not. However, the absence of age related effect is not so surprising as the outcome variable, the mortality rate, is age corrected. The interaction
between the household size and the level of education appears in half of the models, whereas appropriate main effects are not. This interaction could be a proxy to other effects that were not measured. Interactions involving the average precipitation appear less than other predictors. The interaction with humidity usually appears without the main effect of precipitation. Nevertheless, both interactions should take into account when constructing a prediction model for the mortality rate. The next step, that we omit here for brevity, would probably be to look at the sample correlation matrix of the predictors in Table 4.

6 Discussion

Model selection consistency is an ambitious goal to achieve when dealing with high-dimensional data. A “class of minimal models” was defined to be a set of models that should be considered as candidates for prediction of the outcome variable. A search algorithm to identify these models was developed using simulated annealing method and using a “score” for each variable. Under suitable conditions, that are outlined in Theorem 4.1, the algorithm passes through the relevant models.

The score for each predictor was later discussed and a one possible choice was suggested. The suggested procedure uses the Lasso, the Elastic Net and a reduced penalty Lasso. These score are not necessarily the best to use but we claim that they are sensible. Other scoring methods may achieve better results. On the other hand, the scores we use here may be used for other purposes. Theoretical justification for using the Elastic Net to unveil predictors the Lasso might have missed was also presented. While this was not the center of the paper, Theorem 4.3 gave a mathematical justification for the common saying that the Elastic Net is better to use when dealing with correlated variables.

A simulation study was used to demonstrate the capability of the search algorithm to detect relevant models. As illustrated using real data examples, a class of minimal models can be used to derive conclusions regarding the problem at hand. This is rarely the case that a researcher believes a one true model exists, especially in the $p \gg n$ regime. Therefore, we suggest to abandon the search for this “holy grail”, and to analyze the class of minimal models instead.

Appendix

A Proofs

A.1 Proof of Theorem 4.1

We start with the following lemma.

Lemma 1 Assume $Y = \mu + \epsilon$ and assume also (A1)-(A4). Let $S_k = \{S : |S| = k, \hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T Y\}$ be the set of all models with $k$ variables, such
Therefore, \( \hat{\beta}_S \), the LS estimate, is unique. Denote \( S_j^* = S \cup \{ j \}, j \notin S \) for a model that includes \( S \) and additional variable \( j \) not in \( S \). We have

\[
\max_{S \in S_k, 1 \leq j \leq p} \frac{1}{n} \epsilon^T (X_{S_j^*} \hat{\beta}_{S_j^*} - X_S \hat{\beta}_S) = o_p(n)
\]

Proof. Let \( \xi_j \) be the vector of coefficients obtained by regressing \( x_j \), the \( j^{th} \) column in \( X \), on \( X_S \) and let \( P_j \) be the projection operator on the subspace spanned by the part of \( x_j \) which is orthogonal to the subspace spanned by \( X_S \). That is,

\[
P_j = \frac{(x_j - X_S \xi_j)(x_j - X_S \xi_j)^T}{||x_j - X_S \xi_j||^2}.
\]

Let \( \hat{\beta}_{S_j^*} \) be the coefficient estimate of \( x_j \) in model \( S_j^* \), and let \( \hat{\beta}_{S_j^*}^{-1} \) be the coefficient estimates of the variables in \( S \) but for the model \( S_j^* \). Since \( (x_j - X_S \xi_j) \) is orthogonal to the subspace spanned by the columns of \( X_S \) we have

\[
X_{S_j^*} \hat{\beta}_{S_j^*} = x_j \hat{\beta}_{S_j^*}^{||} + X_S \hat{\beta}_{S_j^*}^{-1}
\]

\[
= (x_j - X_S \xi_j) \hat{\beta}_{S_j^*}^{||} + X_S (\hat{\beta}_{S_j^*}^{||} + \xi_j \hat{\beta}_{S_j^*}^{-1})
\]

\[
= (x_j - X_S \xi_j) \hat{\beta}_{S_j^*}^{||} + X_S \hat{\beta}_{S_j^*}^{-1}
\]

\[
= P_j y + X_S \hat{\beta}_S.
\]

Therefore,

\[
\epsilon^T (X_{S_j^*} \hat{\beta}_{S_j^*} - X_S \hat{\beta}_S) = \epsilon^T P_j \mu + \epsilon^T P_j \epsilon.
\]

Now, since \( ||P_j \mu||^2 \leq ||\mu||^2 = O(n) \), we get that for all \( j \), \( \epsilon^T P_j \mu = O_p(\sqrt{n}) \).

Next, denote \( Z_j = P_j \epsilon \), and observe that \( Z_j \sim N(0, \sigma^2) \). We have for any \( a > 0 \)

\[
P\left( \max_{S \in S_k, 1 \leq j \leq p} \frac{1}{n} \epsilon^T P_j \epsilon \geq a \right) = P\left( \max_{S \in S_k, 1 \leq j \leq p} |Z_j| \geq \sqrt{\frac{a n}{\sigma^2}} \right) \leq \sigma \sqrt{\frac{2(k+1) \log p + o(1)}{a n}}
\]

where the last inequality follows since the \( Z_j \)'s are normally distributed and since the approximate size of the set \( \{ S_k \} \times \{1,...,p\} \) is \( p^{k+1} \). Now, since \( p = n^a \) and \( k = o(n/\log n) \) we get that

\[
P\left( \max_{S \in S_k, 1 \leq j \leq p} \frac{1}{n} \epsilon^T P_j \epsilon \geq a \right) = o(1)
\]

and we are done. \( \square \)

We can now move to the proof of Theorem 4.1. For simplicity, the notation of \( i \) as the iteration number for the current temperature \( t \) is suppressed. Note that it is enough to only consider models such that \( S \cap \bar{S} = \emptyset \) and to consider \( m = s_0 \). Denote \( Q_t(S,g,j) \) for the probability of a move in the direction of \( \bar{S} \) in the next iteration, that is, the probability of choosing a variable \( j \in S \cap \bar{S}^c \).
and replace it with a variable \( g \in S' \cap \bar{S} \). Denote \( S' = \{ S/\{j\} \} \cup \{ g \} \) for this new model. We have

\[
Q_t(S, g, j) = P(S \rightarrow S') \min \left[ 1, \exp \left( \frac{||Y - X_S\hat{\beta}_S||^2 - ||Y - X_{S'}\hat{\beta}_{S'}||^2}{t} \right) \right] \frac{P(S' \rightarrow S)}{P(S \rightarrow S')} \tag{13}
\]

where \( P(S \rightarrow S') \) is the probability of suggesting \( S' \), given current model is \( S \). Now, since \( \gamma_{\min} \geq c_\gamma \) and since the maximal value in \( \gamma \) equals to one by definition, we have for all \( S \subseteq A \gamma \)

\[
c_\gamma (h_\gamma - s_0) \leq \sum_{u \not\in S} \gamma_u \leq h_\gamma - s_0
\]

\[
s_0 \leq \sum_{v \in S} \frac{1}{\gamma_v} \leq \frac{s_0}{c_\gamma}.
\tag{14}
\]

Now, by substituting (14) into (8)-(10) we get

\[
P(S \rightarrow S') = \frac{\gamma g \sum_{u \not\in S} \gamma_u}{\gamma_0 \sum_{v \in S} \frac{1}{\gamma_v}} \geq \frac{c_\gamma^2}{s_0 (h_\gamma - s_0)},
\tag{15}
\]

\[
P(S' \rightarrow S) = \frac{\gamma_0^2 \sum_{u \not\in S'} \gamma_u}{\gamma_0 \sum_{v \in S'} \frac{1}{\gamma_v}} \geq \frac{c_\gamma^4}{s_0 (h_\gamma - s_0)}.
\]

Next, we have

\[
\frac{1}{n} ||Y - X_S\hat{\beta}_S||^2 - \frac{1}{n} ||Y - X_{S'}\hat{\beta}_{S'}||^2 = \frac{1}{n} \left( (Y - X_{S'}\hat{\beta}_{S'})^T - (Y - X_{S}\hat{\beta}_{S})^T \right) (X_{S'}\hat{\beta}_{S'} - X_S\hat{\beta}_S)
\]

\[
= \frac{1}{n} \left( (Y - X_{S'}\hat{\beta}_{S'})^T - (Y - X_{S}\hat{\beta}_{S})^T \right) \left( X_{S'}\hat{\beta}_{S'} - X_S\hat{\beta}_S \right)
\]

\[
= \frac{1}{n} \hat{\mu}^T (X_{S'}\hat{\beta}_{S'} - X_S\hat{\beta}_S) + \frac{1}{n} \epsilon^T (X_{S'}\hat{\beta}_{S'} - X_S\hat{\beta}_S)
\]

\[
= \frac{1}{n} \hat{\mu}^T (X_{S'}\hat{\beta}_{S'} - X_S\hat{\beta}_S) + \Delta_n(S, S') \tag{16}
\]

where the second equality is due to \( \hat{\beta}_S \) and \( \hat{\beta}_{S'} \) being LS estimators. We get that an estimator in linear model achieves better (lower) sample MSE, if the correlation of the predication using this estimator with \( Y \) is larger. Now, denote \( S'' = S' \cup S \). We have

\[
\Delta_n(S, S') = \frac{1}{n} \epsilon^T \left( (X_{S''}\hat{\beta}_{S''} - X_S\hat{\beta}_S) - (X_{S'}\hat{\beta}_{S'} - X_S\hat{\beta}_S) \right)
\]

24
and if we apply Lemma 1 twice we get that $\Delta_n(S, S') = o_p(1)$. Now, regarding the first term in $(16)$,
\[
\frac{1}{n} \mu^T (X_{S'} \hat{\beta}_{S'} - X_S \hat{\beta}_S) = \frac{1}{n} \mu^T (\mathcal{P}_{S'} y - \mathcal{P}_S y)
\]
\[
= \frac{1}{n} \left( \|\mathcal{P}_{S'} \mu\|_2^2 - \|\mathcal{P}_S \mu\|_2^2 \right) + \Delta'_n(S, S')
\]
where $\Delta'_n(S, S') = \frac{1}{n} \mu^T [\mathcal{P}_{S'} \epsilon - \mathcal{P}_S \epsilon]$. The content of the proof of Lemma 1 implies that $\Delta'_n(S, S') = o_p(1)$. Now, by $(16)$ and $(17)$ and since Assumption (B1) holds for $t_0$ we get that for large enough $n$
\[
\frac{1}{n} \left( \|Y - X_{S'} \hat{\beta}_{S'}\|^2 - \|Y - X_S \hat{\beta}_S\|^2 \right) \geq 4t \log c_y.
\]
Now, by substituting $(15)$ and $(18)$ into $(13)$ we get that for large enough $n$,
\[
Q_{t_0}(S, g, j) \geq \frac{c^2_2}{s_0(h - s_0)}
\]
for all $S \neq \tilde{S}$, $j \in S \cap S^c$ and $g \in S^c \cap S$. $(11)$ follows from this immediately since for any integer $m$ and for all $S \neq \tilde{S}$,
\[
P^m_{t_0}(S'|S) \geq \min_{\{S: S \cap S^c \neq \emptyset\}} \left[ Q_t(S, g, j) \right]^{s_0} \geq \left[ \frac{c^2_2}{s_0(h - s_0)} \right]^{s_0}.
\]

A.2 Proof of Proposition 4.2
Recall that the Elastic Net estimator $\hat{\beta}^{EN}$ minimize
\[
||Y - X\beta||_2^2 + \lambda_1 |\beta| + \lambda_2 ||\beta||_2^2
\]
Now, WLOG assume that $\hat{\beta}^{EN}$ is a solution such that $\hat{\beta}^{EN}_1 > 0$. For convenience, we omit the “EN” superscript from now on (i.e., $\beta = \hat{\beta}^{EN}$). Define the subspace
\[
B := \{ \beta : \forall i \neq 1, 2 \beta_i = \hat{\beta}_i, \beta_1 = \tau \hat{\beta}_1, \beta_2 = (1 - \tau) \hat{\beta}_2 \}.
\]
WLOG, write down $X$ as $X = (X_{(12)} \quad X_{(1-12)})$ where $X_{(12)} = (x_1 \quad x_2)$ are the first two columns of $X$ and $X_{(1-12)}$ are the rest of its columns. Similarly, we have $\beta^T = (\beta_{(12)} \quad \beta_{(1-12)})$ where $\beta_{(12)}$ is the first two entries in the vector $\beta$ and $\beta_{(1-12)}$ is the rest of the vector. Define $\tilde{Y} = Y - X_{(1-12)} \beta_{(1-12)}$. We can rewrite $(19)$ as
\[
||\tilde{Y} - X_{(12)} \beta_{(12)}||^2 + \lambda_1 (|\beta_{(12)}| + |\beta_{(1-12)}|) + \lambda_2 (||\beta_{(12)}||_2^2 + ||\beta_{(1-12)}||_2^2)
\]
\[
\text{If the minimum of } (21), \text{ on } B, \text{ is achieved at } 0 < \tau^* < 1 \text{ then } \hat{\beta}_2 \text{ must be non zero. Minimizing } (21) \text{ on } B \text{ is essentially minimizing}
\]
\[
-2\tilde{Y}^T X_{(12)} \beta_{(12)} + ||X_{(12)} \beta_{(12)}||_2^2 + \lambda_2 \beta_{(12)} \beta_{(1-12)}^T
\]
on \( B \). Now, by the definition of \( B \) in (20) and using simple algebra we get that (22) equals to

\[
2 \left[ \hat{\beta}_1 \hat{Y}^T (\tau(x_2 - x_1) - x_2) - \hat{\beta}_1^2 \tau(1 - \tau)(1 - \rho) + \lambda_2 \hat{\beta}_1^2 \left( \frac{1}{2} - \tau(1 - \tau) \right) \right].
\]

This is a quadratic function of \( \tau \), and by equating its derivative to zero we get that

\[
\tau^* = \frac{1}{2} - \frac{\hat{Y}^T(x_2 - x_1)}{2\hat{\beta}_1(\lambda_2 + 1 - \rho)}
\]

is the minimizer of (19) (the coefficient of the quadratic term is positive). Note that for \( x_2 = x_1 \) we get the expected \( \tau^* = \frac{1}{2} \) solution. Note also that this reveals no information regarding the Lasso where \( \lambda_2 = 0 \). Next, we get that \( 0 < \tau^* < 1 \) if

\[
\left| \frac{\hat{Y}^T(x_2 - x_1)}{\hat{\beta}_1(\lambda_2 + 1 - \rho)} \right| < 1.
\]

(23)

Since \( ||x_2 - x_1||^2_2 = 2(1 - \rho) \) we have

\[
|\hat{Y}^T(x_2 - x_1)| \leq \sum_{i=1}^{n} |\hat{Y}_i||x_{2i} - x_{1i}| \leq ||\hat{Y}||_2 \sqrt{2(1 - \rho)},
\]

using the triangle inequality and then Cauchy-Schwartz inequality. It is assumed that \( \hat{\beta}_1 \geq c_\beta > 0 \) and it is known that \( ||\hat{Y}||_2 \leq ||Y||_2 \). Therefore, we may rewrite (23) as

\[
\frac{\sqrt{2}||Y||_2 \sqrt{1 - \rho}}{c_\beta(\lambda_2 + 1 - \rho)} < 1.
\]

Now, Denote \( t = \sqrt{1 - \rho} \), \( u = \frac{||Y||_2}{c_\beta} \), we have

\[
t^2 - \sqrt{2}ut + \lambda_2 > 0.
\]

For \( \lambda_2 > \frac{1}{2}u^2 \), we get the result we want for all \( \rho \)'s. For \( \lambda_2 < \frac{u^2}{2} \) we have

\[
\sqrt{1 - \rho} > \frac{1}{\sqrt{2}}(u + \sqrt{u^2 - 2\lambda_2}),
\]

(24)

\[
\sqrt{1 - \rho} < \frac{1}{\sqrt{2}}(u - \sqrt{u^2 - 2\lambda_2}).
\]

(25)

The RHS of (24) is larger than 1 if \( \lambda_2 < \sqrt{2}u - 1 \). That is, there is no suitable \( \rho \) for this case. The RHS of (25) is always positive, and for the same condition \( \lambda_2 < \sqrt{2}u - 1 \), it also meaningful, i.e., \( (u - \sqrt{u^2 - 2\lambda_2}) < \sqrt{2} \) and in terms of \( \rho \),

\[
\rho > 1 - \frac{1}{2}(u - \sqrt{u^2 - 2\lambda_2})^2
\]

or alternatively,

\[
\rho > 1 - \frac{u^2}{2} \left( 1 - \sqrt{1 - \frac{2\lambda_2}{u}} \right)^2.
\]
and by Taylor expansion for \( \frac{2\lambda_2}{u} \) we get
\[
\rho > 1 - \frac{\lambda_2^2}{2u^2}
\]
\[ \square \]

### A.3 Proof of Theorem 4.3

The proof is similar to the proof of Proposition 4.2. Let \( \hat{\beta} = \hat{\beta}^{EN} \) be the Elastic Net estimator and denote \( \hat{\beta}_M \) for the values in \( \beta \) corresponding to the set of predictors \( M \). We can partition the set of potential predictors \( \{1, 2, \ldots, p\} \) to four disjoint subsets: \( M'^{-} \); \( M_1 \cap M_2' \); \( M_1' \cap M_2 \) and \( M_1 \cap M_2 \). We replace (20) with
\[
B := \{ \beta : \beta_{M'^{-}} = \hat{\beta}_{M'^{-}}, \beta_{M_1 \cap M_2} = \hat{\beta}_{M_1 \cap M_2}, \beta_{M_1' \cap M_2} = (1 - \tau)\Theta' \hat{\beta}_{M_1 \cap M_2}' \}
\]
where \( \beta_M \) is defined as the values in \( \hat{\beta}_M \) corresponding to the set \( M \) and \( \Theta' \) is the matrix of coefficients obtained from regressing \( X_{M_1 \cap M_2} \) on \( X_{M_1'} \cap M_2 \). We define \( \Theta \) to be an augmented version of \( \Theta' \), which we obtain by regressing \( X_{M_1} \) on \( X_{M_2} \). That is,
\[
X_{M_2} \Theta = P_{M_2} X_{M_1}
\]
Note that on \( B \),
\[
X \beta = X_{M_1} \hat{\beta}_{M_1} + \tau X_{M_1} \hat{\beta}_{M_2} + (1 - \tau) X_{M_2} \Theta \hat{\beta}_{M_1}
\]
Recalling that \( \tilde{Y} = Y - X_{M_1} \hat{\beta}_{M_1} \), minimizing [4] on \( B \) is equivalent to minimize
\[
||\tilde{Y} - \tau X_{M_1} \hat{\beta}_{M_1} - (1 - \tau) X_{M_2} \Theta \hat{\beta}_{M_1}||_2^2 + \lambda_1 ||\hat{\beta}_{M_1}||_1 + (1 - \tau)||\Theta \hat{\beta}_{M_1}||_1
\]
\[+ \lambda_2 [\tau^2 ||\hat{\beta}_{M_1}||_2^2 + (1 - \tau)^2 ||\Theta \hat{\beta}_{M_1}||_2^2]
\]
as a function of \( \tau \). Using a first-order condition and substituting (28) we find that (29) is minimized for
\[
\tau^* = \frac{-(\tilde{Y} - P_{M_2} X_{M_1} \hat{\beta}_{M_1})^T (I - P_{M_2}) X_{M_1} \hat{\beta}_{M_1} + \frac{\lambda_1}{2} (||\hat{\beta}_{M_1}||_1 - ||\Theta \hat{\beta}_{M_1}||_1) - \lambda_2 ||\Theta \hat{\beta}_{M_1}||_2^2}{||I - P_{M_2}|| X_{M_1} \hat{\beta}_{M_1}||_2^2 - \lambda_2 ||\beta_{M_1}||_2^2 - \lambda_2 ||\Theta \hat{\beta}_{M_1}||_2^2}
\]
where \( I \) is the identity matrix. Before we continue, note that if \( X_2 = X_1 \) then \( \Theta \) is the identity matrix and \( P_{M_2} X_{M_1} = X_1 \). Substituting these facts into (30), we get that \( \tau^* = \frac{1}{2} \) as one my expects. Same result is obtained for the case \( M_2 \subseteq M_1 \).
As it can be seen in (26), the coordinates of $\hat{\beta}_{M_2}$ are all different than zero if $\tau^* < 1$. Now, since $\mathcal{P}_{M_2}(I - \mathcal{P}_{M_2}) = 0$ we get that $\tau^* < 1$ if

$$-Y^T(I - \mathcal{P}_{M_2})X_{M_1}\hat{\beta}_{M_1} + \|(I - \mathcal{P}_{M_2})X_{M_1}\hat{\beta}_{M_1}\|_2^2 - \frac{\lambda_1}{2}||\Theta\hat{\beta}_{M_1}||_1$$

$$> -\frac{\lambda_1}{2}||\hat{\beta}_{M_1}||_1 - \lambda_2||\hat{\beta}_{M_1}||_2^2$$

which is certainly true if

$$Y^T\mathcal{P}_{M_2}X_{M_1}\hat{\beta}_{M_1} - \frac{\lambda_1}{2}||\Theta\hat{\beta}_{M_1}||_1 > -\frac{\lambda_1}{2}||\hat{\beta}_{M_1}||_1 - \lambda_2||\hat{\beta}_{M_1}||_2^2 + Y^TX_{M_1}\hat{\beta}_{M_1} \quad (31)$$

which is true if the condition in (12) is fulfilled for the appropriate $c_1$. 

\[\square\]

B Supplementary tables for Section 5.2

Table 5: The 112 models selected for the riboflavin data. Each row is a model, the numbers are the column number (the gene) in $X$.

| Model | 1278 | 1279 | 4003 | 1516 | 1312 | 1278 | 4003 | 1303 | 4003 | 1279 | 4003 | 1278 | 4006 | 1279 | 4004 | 1279 | 4003 | 69 | 2564 | 4003 | 73 | 2564 | 4003 | 144 | 2564 | 4003 | 69 | 2564 | 4004 | 69 | 2564 | 4006 | 792 | 1478 | 4002 | 792 | 1478 | 4003 | 792 | 1478 | 4004 | 73 | 1279 | 2564 | 4004 | 73 | 1279 | 2564 | 4003 | 144 | 1279 | 2564 | 4004 | 73 | 1849 | 2564 | 4004 | 73 | 1279 | 2564 | 4006 | 144 | 1279 | 2564 | 4003 | 73 | 1279 | 1849 | 2564 | 4003 | 69 | 1849 | 2564 | 3226 | 4003 |
|   |   |   |   |   |   |
|---|---|---|---|---|---|
| 27 | 69 | 1425 | 1640 | 2564 | 4003 |
| 28 | 69 | 1849 | 2564 | 3226 | 4004 |
| 29 | 69 | 1425 | 1640 | 2564 | 4004 |
| 30 | 144 | 792 | 1849 | 2564 | 4003 |
| 31 | 144 | 1849 | 2564 | 3226 | 4003 |
| 32 | 73 | 974 | 1279 | 2564 | 4003 |
| 33 | 144 | 1278 | 1425 | 2564 | 4003 |
| 34 | 73 | 792 | 2116 | 2564 | 4004 |
| 35 | 73 | 1278 | 1849 | 2564 | 4004 |
| 36 | 144 | 1278 | 1849 | 2564 | 4003 |
| 37 | 73 | 1279 | 1425 | 2564 | 4004 |
| 38 | 144 | 792 | 1849 | 2027 | 2564 | 4004 |
| 39 | 73 | 974 | 1278 | 1849 | 2564 | 4003 |
| 40 | 69 | 415 | 1849 | 2564 | 3226 | 4004 |
| 41 | 73 | 792 | 974 | 2116 | 2564 | 4003 |
| 42 | 73 | 1279 | 1640 | 1849 | 2564 | 4004 |
| 43 | 69 | 315 | 792 | 1849 | 2564 | 4004 |
| 44 | 73 | 792 | 1849 | 2027 | 2564 | 4004 |
| 45 | 73 | 792 | 1303 | 2116 | 2564 | 4003 |
| 46 | 69 | 792 | 1849 | 2027 | 2564 | 4003 |
| 47 | 69 | 792 | 1282 | 1849 | 2564 | 4003 |
| 48 | 69 | 792 | 1131 | 1849 | 2564 | 4003 |
| 49 | 73 | 1131 | 1278 | 1524 | 2564 | 4006 |
| 50 | 144 | 1131 | 1303 | 1524 | 2564 | 4006 |
| 51 | 73 | 792 | 1528 | 1849 | 2564 | 4003 |
| 52 | 73 | 792 | 1294 | 2116 | 2564 | 4003 |
| 53 | 69 | 1131 | 1278 | 1524 | 1762 | 2564 | 4006 |
| 54 | 144 | 1279 | 1762 | 1820 | 2027 | 2564 | 4004 |
| 55 | 69 | 1279 | 1425 | 1640 | 1820 | 2564 | 4006 |
| 56 | 144 | 792 | 1312 | 1849 | 2027 | 2564 | 4004 |
| 57 | 73 | 1278 | 1762 | 1820 | 1857 | 2564 | 4003 |
| 58 | 73 | 1131 | 1279 | 1524 | 1528 | 2564 | 4003 |
| 59 | 69 | 792 | 1303 | 1849 | 2484 | 2564 | 4003 |
| 60 | 69 | 792 | 1639 | 1849 | 2027 | 2564 | 4003 |
| 61 | 73 | 315 | 792 | 1278 | 1524 | 2564 | 4004 |
| 62 | 69 | 315 | 1425 | 1524 | 1640 | 2564 | 4004 |
| 63 | 73 | 1131 | 1279 | 1857 | 2116 | 2564 | 4004 |
| 64 | 144 | 1131 | 1279 | 1857 | 2116 | 2564 | 4004 |
| 65 | 144 | 1101 | 1131 | 1279 | 1762 | 2564 | 4004 |
| 66 | 69 | 792 | 1131 | 1849 | 2564 | 3514 | 4004 |
| 67 | 73 | 792 | 1279 | 1478 | 2027 | 2564 | 4002 |
| 68 | 73 | 792 | 1131 | 1279 | 1312 | 2116 | 2564 | 4004 |
| 69 | 73 | 315 | 792 | 1279 | 1312 | 2116 | 2564 | 4004 |
| 70 | 73 | 315 | 792 | 1279 | 1503 | 2116 | 2564 | 4004 |
| 71 | 69 | 792 | 1131 | 1279 | 2116 | 2564 | 3288 | 4006 |
| 72 | 73 | 315 | 1279 | 1762 | 1849 | 2564 | 3288 | 4004 |
|   | 144  | 974  | 1131 | 1279 | 1524 | 2564 | 3514 | 4003 |
|---|------|------|------|------|------|------|------|------|
| 74| 144  | 974  | 1131 | 1279 | 1425 | 1524 | 2564 | 4003 |
| 75| 69   | 792  | 859  | 1131 | 1279 | 2116 | 2564 | 4004 |
| 76| 73   | 792  | 1279 | 1849 | 2484 | 2564 | 4004 | 4006 |
| 77| 73   | 974  | 1101 | 1131 | 1279 | 2564 | 3105 | 4004 |
| 78| 73   | 792  | 1131 | 1312 | 2116 | 2242 | 2564 | 4004 |
| 79| 73   | 792  | 974  | 1131 | 1364 | 2116 | 2564 | 4004 |
| 80| 73   | 792  | 1131 | 1312 | 1639 | 2116 | 2564 | 4004 |
| 81| 73   | 792  | 1131 | 1364 | 2116 | 2242 | 2564 | 4004 |
| 82| 73   | 792  | 1131 | 1312 | 2116 | 2564 | 3905 | 4004 |
| 83| 144  | 1131 | 1279 | 1524 | 1528 | 2484 | 2564 | 4004 |
| 84| 73   | 974  | 1131 | 1279 | 1524 | 2242 | 2564 | 3514 | 4006 |
| 85| 73   | 974  | 1131 | 1279 | 1524 | 2242 | 2564 | 3465 | 4006 |
| 86| 73   | 244  | 792  | 1131 | 1278 | 2116 | 2564 | 3104 | 4006 |
| 87| 73   | 792  | 1131 | 1278 | 1297 | 2116 | 2564 | 3104 | 4006 |
| 88| 144  | 1101 | 1279 | 1425 | 1640 | 2116 | 2564 | 4004 |
| 89| 69   | 859  | 1101 | 1640 | 1762 | 2484 | 2564 | 3226 | 4003 |
| 90| 73   | 315  | 792  | 1303 | 1849 | 2564 | 4004 | 4006 | 4045 |
| 91| 73   | 144  | 315  | 792  | 1279 | 1849 | 2462 | 2564 | 4004 |
| 92| 73   | 315  | 792  | 1303 | 1849 | 2564 | 4004 | 4045 | 4075 |
| 93| 73   | 827  | 1131 | 1279 | 1639 | 2242 | 2564 | 3465 | 4003 |
| 94| 69   | 1312 | 1425 | 1640 | 1762 | 2116 | 2564 | 3104 | 4004 |
| 95| 69   | 1312 | 1425 | 1528 | 1640 | 1762 | 2116 | 2564 | 4004 |
| 96| 73   | 624  | 792  | 1131 | 1278 | 1849 | 1855 | 2564 | 4004 |
| 97| 144  | 827  | 1131 | 1279 | 1639 | 2242 | 2564 | 3465 | 4003 |
| 98| 73   | 792  | 1131 | 1279 | 2027 | 2116 | 2564 | 3104 | 3288 | 4004 |
| 99| 73   | 974  | 1131 | 1279 | 1364 | 1524 | 2027 | 2564 | 3104 | 4003 |
| 100| 73   | 859  | 974  | 1131 | 1279 | 1364 | 1524 | 2462 | 2564 | 4003 |
| 101| 73   | 624  | 792  | 1131 | 1279 | 1849 | 2116 | 2564 | 3226 | 4004 |
| 102| 73   | 1303 | 1524 | 1762 | 2027 | 2484 | 2564 | 3905 | 4004 | 4075 |
| 103| 69   | 974  | 1425 | 1524 | 1640 | 2116 | 2484 | 2564 | 3288 | 4004 |
| 104| 69   | 974  | 1278 | 1425 | 1524 | 1640 | 2484 | 2564 | 3288 | 4004 |
| 105| 73   | 315  | 1279 | 1294 | 1762 | 2027 | 2462 | 2564 | 4003 | 4075 |
| 106| 69   | 974  | 1278 | 1425 | 1524 | 1640 | 1857 | 2484 | 2564 | 3288 | 4004 |
| 107| 69   | 974  | 1425 | 1524 | 1640 | 1857 | 2484 | 2564 | 3288 | 4004 |
| 108| 73   | 859  | 1131 | 1278 | 1297 | 1524 | 2462 | 2564 | 3465 | 4003 |
| 109| 73   | 1278 | 1425 | 1524 | 1639 | 2484 | 2564 | 3465 | 4003 | 4045 |
| 110| 73   | 792  | 1131 | 1279 | 1478 | 2116 | 2484 | 2564 | 3465 | 4006 |
| 111| 73   | 859  | 1131 | 1278 | 1503 | 1524 | 2462 | 2484 | 2564 | 4003 |
| 112| 73   | 415  | 859  | 1131 | 1278 | 1503 | 1524 | 2484 | 2564 | 4003 |
Table 6: Potential predictors for mortality rate in Section 5.2.2

| Predictor | Description |
|-----------|-------------|
| prec      | Mean annual precipitation in inches |
| jant      | Mean January temperature in degrees F |
| jult      | Mean July temperature in degrees F |
| age65     | Percentage of population aged 65 or older |
| pphs      | Population per household |
| educ      | Median school years completed by those over 22 |
| facl      | Percentage of housing units which are sound and with all facilities |
| dens      | Population per square mile in urbanized areas |
| nwht      | Percentage of non-white population in urbanized areas |
| wtcl      | Percentage of employed in white collar occupations |
| linc      | Percentage of families with income ≥ 3000 dollars in urbanized areas |
| HC        | Relative pollution potential of hydrocarbon |
| NOx       | Relative pollution potential of nitric oxides |
| SUL       | Relative pollution potential of sulphur dioxide |
| hum       | Annual average percentage of relative humidity at 1pm |

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6), 716–723.

Bickel, P. J. & Cai, M. (2012). Discussion of sara van de geer: Generic chaining and the $\ell_1$ penalty. *Journal of Statistical Planning and Inference*.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4), 1705–1732.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2010). Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown* (pp. 56–69).

Bien, J., Taylor, J., Tibshirani, R., et al. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3), 1111–1141.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.

Brooks, S., Friel, N., & King, R. (2003). Classical model selection via simulated annealing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2), 503–520.

Brooks, S. & Morgan, B. (1995). Optimization using simulated annealing. *The Statistician*, 241–257.
Bühlmann, P., Kalisch, M., & Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application, 1*, 255–278.

Bunea, F., Tsybakov, A., & Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics, 1*, 169–194.

Bunea, F., Tsybakov, A. B., & Wegkamp, M. H. (2006). Aggregation and sparsity via 1 penalized least squares. In *Learning theory* (pp. 379–391).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1.

Greenshtein, E. & Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli, 10*(6), 971–988.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika, 57*(1), 97–109.

Hjort, N. L. & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association, 98*(464), 879–899.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science, 382–401.*

Huang, J., Ma, S., & Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica, 18*(4), 1603.

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

Kirkpatrick, S., Jr., D. G., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

McDonald, G. C. & Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics, 15*(3), 463–481.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics, 34*(3), 1436–1462.

Meinshausen, N. & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 417–473.

Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association, 104*(488).
Meinshausen, N. & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 246–270.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21, 1087.

Rigollet, P. & Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statistical Science*, 27(4), 558–575.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.

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

Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *Information Theory, IEEE Transactions on*, 50(10), 2231–2242.

Wang, H., Li, B., & Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3), 671–683.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49–67.

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

Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10(3).

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.