Journal of Machine Learning Research 18 (2017) 1-29 Submitted 4/16; Revised 1/17; Published 4/17

Identifying a Minimal Class of Models for High–dimensional Data

Daniel Nevo DANIELNEVO@GMAIL.COM
Department of Statistics
The Hebrew University of Jerusalem
Mt. Scopus, Jerusalem, Israel
and
Current address: Departments of Biostatistics and Epidemiology
Harvard T.H. Chan School of Public Health
Boston, MA 02115, USA

Ya'acov Ritov YAACOV.RITOV@GMAIL.COM
Department of Statistics
The Hebrew University of Jerusalem
Mt. Scopus, Jerusalem, Israel
and
Department of Statistics
University of Michigan
Ann Arbor, MI 48109–1107, USA

Editor: Nicolai Meinshausen

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 a minimal class of models. The minimal class of 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 minimal class of models is demonstrated in the analysis of two data sets.

Keywords. Model Selection; High–dimensional Data; Lasso; Elastic Net; Simulated Annealing

1. Introduction

High–dimensional statistical problems have been arising as a result of the vast amount of data gathered today. A more specific problem is that estimation of the usual linear regression coefficients vector cannot be performed when the number of predictors exceeds the number of observations. 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 have 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 an ℓ_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 thoroughly researched in the past 15 years. For the high–dimensional problem, prediction rates were established in various manners (Greenshtein and Ritov, 2004; Bunea et al., 2006; Bickel et al., 2009; Bunea et al., 2007; Meinshausen and Yu, 2009). The capability of the lasso to choose the correct model depends on the true coefficient vector and the matrix of the predictors (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Zhang and Huang, 2008). However, the underlying assumptions are typically rather restrictive, and cannot be checked in practice.

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 in the classical setup, with more observations than predictors, there is no model selection consistent estimator unless further assumptions are fulfilled. This leads us to present a different objective. Instead of searching for a single "true" model, we aim to present a number of possible models a researcher should look at. Our goal, therefore, is to find potentially good prediction models. In short, we suggest to find the best models for each small model size. Then, by looking at these models one may reach interesting conclusions regarding the underlying problem. Some of these, as we demonstrate in applications, can be concluded using statistical reasoning, but most of these should be reasoned by a subject matter expert.

In order to find these models, we implement a search algorithm that uses simulated annealing (Kirkpatrick et al., 1983). The suggested algorithm is provided with a "score" for each predictor. We suggest to get these scores using a multi–step procedure that implements both the lasso and the elastic net (Zou and Hastie, 2005) (and then the lasso again). Multi–step procedures in the high–dimensional setting have drawn some attention and were demonstrated to be better than the standard lasso (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 describes a search algorithm for relevant models, and gives motivation for the sequential use of the lasso and the elastic net when calculating scores for the predictors. Section 4 investigates the performance of the suggested search algorithm in simulation studies and then Section 5 illustrates data analysis using a minimal class of models in two examples. Section 6 suggests a short discussion. Technical proofs and supplementary data are provided in the appendix.

2. Description of the problem

We start with notations. First, denote $||v||_q := (\sum v_j^q)^{1/q},\ q > 0$ for the ℓ_q (*pseudo*) norm of any vector v and $\|v\|_0 = \lim_{q\to 0} ||v||_q$ for its cardinality. The data consist of a matrix of predictors $X_{n\times p} = (X^{(1)}\ X^{(2)}\ ...\ X^{(p)})$ and a response vector $Y_{n\times 1}$. WLOG, X is centered and scaled and Y is centered as well. We are mainly interested in the case $p > n$. The

underling model is $Y = X\beta^0 + \epsilon$ where $\epsilon_{n\times 1}$ is a random error, $E(\epsilon) = 0$, $V(\epsilon) = \sigma^2 I$, I is the identity matrix.

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 $S_0 := \{j : \beta^0 \neq 0\}$ 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 $\hat{\beta}_S^{LS}$ be the usual least squares (LS) estimator corresponding to a model S, that is,

$$\hat{\beta}_S^{LS} = (X_S^T X_S)^{-1} X_S^T Y,$$

provided $X_S^T X_S$ is non singular.

Now, the straightforward approach to estimate S_0 given a model size κ is to consider the following optimization problem:

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

Unfortunately, typically, solving (1) is computationally infeasible. 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 (Tibshirani, 1996), defined as

$$\hat{\beta}^L = \underset{\beta}{\operatorname{argmin}} \Big(\frac{1}{n} ||Y - X\beta||_2^2 + \lambda ||\beta||_1 \Big)$$

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

$$\hat{\beta}_w^L = \underset{\beta}{\operatorname{argmin}} \Big(\frac{1}{n} ||Y - X\beta||_2^2 + \lambda ||w \cdot \beta||_1 \Big) \tag{2}$$

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. Next is the elastic net estimator

$$\hat{\beta}^{EN} = \underset{\beta}{\operatorname{argmin}} \Big(\frac{1}{n} ||Y - X\beta||_2^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2 \Big). \tag{3}$$

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

$$\hat{\beta}^{EN} = \underset{\beta}{\operatorname{argmin}} \Big(\frac{1}{n} ||Y - X\beta||_2^2 + \lambda \big(\alpha ||\beta||_1 + (1 - \alpha) ||\beta||_2^2 \big) \Big). \tag{4}$$

Let $\hat{\beta}_n$ be a sequence of estimators for β^0 and let $\hat{S}_n$ be the sequence of corresponding models. Model selection consistency is commonly defined as

$$\lim_{n\to\infty} P(\hat{S}_n = S_0) = 1. \tag{5}$$

If $p \ll n$ and small, then criteria based methods (e.g., BIC, Schwarz, 1978) are model selection consistent if p is fixed or if suitable conditions are fulfilled, see Wang et al. (2009)

and references therein. However, these methods are rarely computationally feasible for large p. For $p > n$, it turns out that practically strong and unverifiable conditions are needed to achieve (5) for popular regularization based estimators (Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Huang et al., 2008; Jia and Yu, 2010; Tropp, 2004; Zhang, 2009).

In light of these established results, we suggest to pursue a different goal. Instead of finding a single model, we suggest to look for a group of models. Each of these models should include low number of predictors, but it should also be capable of predicting Y well enough. Therefore, $\mathcal{G}_0^n = \mathcal{G}_0^n(\kappa, \eta)$ is called a minimal class of models of size κ and efficiency η if

$$\mathcal{G}_0^n = \Big\{ S : |S| = \kappa \;\&\; E||Y - X_S\beta_S^{0\star}||_2^2 \leq \min_{|S'|=\kappa} E||Y - X_{S'}\beta_{S'}^{0\star}||_2^2\} + \eta \Big\}.$$

where $\beta_S^{0\star}$ minimizes $E||Y - X_S\beta_S||_2^2$. Note that $\mathcal{G}_0^n$ depends on n, similarly to the way the true model has been considered previously (Greenshtein and Ritov, 2004; Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006). In fact, if $p > n$ then, necessarily, p grows with n and, at the least, new potential predictors are added; the size of $\mathcal{G}_0^n$ may also change. Clearly, $\mathcal{G}_0^n$ is unknown. However, at a first sight, it is unclear how to refer to this set. Even when carrying out model selection, one hardly treats the true model as a parameter, even though it can be looked as such. In terms of inference, in simple, low dimensional, models, likelihood ratio tests can be used in some frequentist cases, or posterior probabilities for the Bayesians. More generally, model selection criteria, such as AIC (Akaike, 1974) or BIC, are typically used, without assigning formal statistical tools to address uncertainty. In this sense, the true minimal class of models is some type of a weak equivalence class, with respect to the true model S_0. Given η, it answers the following question. Are there additional models, other than S_0, that can be considered satisfactory? If the answer is yes, which are these models? Existence of alternative models may help researchers to question the strength of certain conclusions. Furthermore, when high–dimensional regression is used as part of a pilot study to determine which predictors should be measured regularly, alternative models may be compared in terms of cost. On the other hand, if there are multiple models explaining the data in a satisfactory manner, is there such a thing as the true model? If not, then conceptual concerns may arise. We further discuss these and related important issues in Section 6.

From a practical point of view, it would be useful to consider instead a sample version of $\mathcal{G}_0^n$, $\mathcal{G}^n(\kappa, \eta)$, defined as

$$\mathcal{G}^n(\kappa, \eta) = \Big\{ S : |S| = \kappa \;\&\; \frac{1}{n}||Y - X_S\hat{\beta}_S^{LS}||_2^2 \leq \min_{|S'|=\kappa} \{\frac{1}{n}||Y - X_{S'}\hat{\beta}_{S'}^{LS}||_2^2\} + \eta \Big\}. \quad (6)$$

One could control how similar the models in $\mathcal{G} = \mathcal{G}^n$ are to each other in terms of prediction, using the tuning parameter η. A reasonable choice is $\eta = c\sigma^2$ with some $c > 0$. If σ^2 is unknown, it could be replaced with an estimate, e.g., using the scaled lasso (Sun and Zhang, 2012). An alternative to $\mathcal{G}$ is to generate the set of models by simply choosing for each κ the M models having the lowest sample mean square error (MSE), for some number M. The LS estimator, $\hat{\beta}_S^{LS}$, minimizes the sample prediction error for any model S with size $s \leq n$. Thus, this estimator is used for each of the considered models.

In practice, one may find $\mathcal{G}$ for a few values of κ, e.g., $\kappa = 1, ..., 10$, and then examines the pooled results, $\bigcup_{j=1}^{k} \mathcal{G}(j, \eta)$. Alternatively, the empirical MSE $n^{-1}||Y - X_S\hat{\beta}_S^{LS}||_2^2$ in the

definition of $\mathcal{G}$ can be replaced with one of the available model selection criteria, e.g., AIC, BIC or lasso. Then, models of varying size can be included in the class. Note that we are interested in situations where there is a fair number of models with a relatively very small number of variables (predictors) out of the available p.

At this point, a natural question is how can we benefit from using a minimal class of models. Examining the models in $\mathcal{G}$ may allow us to derive conclusions regarding the importance of different explanatory variables. We demonstrate this kind of analysis in Section 5 using two real data examples.

A minimal class of models could be also used in conjunction with the available models aggregation procedures. Aggregation of estimates obtained by different models was suggested both for the frequentist (Hjort and Claeskens, 2003), and for the Bayesian, (Hoeting et al., 1999). 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. We, however, are not interested in improving prediction per se, but in identifying good models. Nor are we interested in identifying the best model, since this is not possible or even meaningful in our setup, but in identifying predictors (and models) that are potentially relevant and important.

2.1 Relation to other work

A similar point of view on the relevance of a predictor was given by Bickel and Cai (2012). They considered a predictor to be important if its relative contribution to the predictive power of a set of predictors is high enough. Their next step was to consider only specific type of sets, such that their prediction error is low, yet they do not contain too many variables.

Rigollet and Tsybakov (2012) investigated the question of prediction under minimal conditions. They showed that linear aggregation of estimators is beneficial for high–dimensional regression when assuming sparsity of the number of estimators included in the aggregation. They also showed that choosing exponential weights for the aggregation corresponds to minimizing a specific, yet relevant, penalized problem. Their estimator, however, is computationally impossible and they have little interest in predictors and model identification.

As we describe in Section 3, our suggested search algorithm for candidate models travels through the model space. We choose to use simulated annealing to prevent the algorithm from getting stuck in a local minimum. Some Bayesian model selection procedures move along the model space, usually using a relevant posterior distribution, cf. O'Hara and Sillanpää (2009). We, however, do not assume any prior distribution for the coefficient values. Our use of the algorithm is only as a search mechanism, simply to find as many as possible models included in $\mathcal{G}$. Convergence properties of the classical simulated annealing algorithm are not of interest to our use of it. We are interested in the path generated by the algorithm and not in its final state.

3. A search algorithm

In this section, we suggest an algorithm to find $\mathcal{G}$ for a given κ and η. The problem is that $||Y - X_S\hat{\beta}_S^{LS}||_2^2$ is unknown for all S, and since p is large, even for a relatively small κ,

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 $\mathcal{M}(\kappa)$. $\mathcal{M}$ is a large set of models, but not too large so we can calculate MSEs for all the models within $\mathcal{M}$ in a reasonable computer running time. Once we have $\mathcal{M}$ and the corresponding MSEs, we can form $\mathcal{G}$ by choosing the relevant models out of $\mathcal{M}$.

The remaining question is how to assemble $\mathcal{M}$ for a given κ. Any greedy algorithm is bound to find models that are all very similar. Our purpose is to find models that are similar in their predictive power, but heterogeneous in their structure. For this we propose a simulated annealing algorithm (Kirkpatrick et al., 1983) which we now describe.

3.1 Simulated annealing algorithm

Our approach therefore is to implement a search algorithm which travels between potentially attractive models. We use a simulated annealing algorithm (Kirkpatrick et al., 1983), originally suggested for function optimization. The maximizer of a function $f(\theta)$ is of interest. Let $T = (t_1, t_2, ..., t_R)$ be a decreasing set of positive "temperatures". For every temperature level $t \in T$, iterative steps are carried out, before moving to the next, lower, temperature level. In each step, a random move from the current θ to another $\theta' \neq \theta$ is suggested. The move is then accepted with a probability that depends on $\exp\big[\big(f(\theta') - f(\theta)\big)/t\big]$. Typically, although not necessarily, a Metroplis–Hastings criterion (Metropolis et al., 1953; Hastings, 1970) is used to decide whether to accept the suggested move θ' or to stay at θ. After a predetermined number ,N_t, of such iterations, the algorithm moves to the next $t' < t$ in T, taking the final state in temperature t as the initial state for t'. The motivation for using this algorithm is that for high "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 temperature, the decision to move to a suggested point is based almost solely on the criterion of improvement in the target function value. The name of the algorithm and its motivation come from annealing in metallurgy (or glass processing), where a strained piece of metal is heated, so that a reorganization of its atoms is possible, and then it colds off slowly so the atoms can settle down in low energy position. See Brooks and Morgan (1995) for a general review of simulated annealing in the context of statistical problems.

In our case, the parameter of interest is the model S and the objective function is

$$f(S) = -\frac{1}{n}||Y - X_S\hat{\beta}_S^{LS}||_2^2.$$

We now describe the proposed algorithm in more detail. We use simulated annealing with Metropolis–Hastings acceptance criterion as a search mechanism for good models. That is, we are not looking for the settling point of the algorithm; instead, we are following its path, hoping that much of it will be in neighborhood of good models.

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

at S_t^i and $\hat{\beta}_t^i$. We now define how S_t^{i+} is suggested and what is the probability of accepting this move.

For each S_t^i, we suggest S_t^{i+} by a minor change, i.e., we take one predictor out and we add another instead, and then obtain $\hat{\beta}_t^{i+}$ by standard linear regression. Assume that for every variable $j \in (1, ..., p)$, we have a score γ_j, such that higher value of γ_j reflects that the variable j should be included in a model, comparing with other possible variables. WLOG, assume $0 \leq \gamma_j \leq 1$ for all j. We choose a variable $r^* \in S_t^i$ and take it out with the probability function

$$p_{i,r}^{out} = \frac{{\gamma_r}^{-1}}{\sum\limits_{u \in S_t^i} {\gamma_u}^{-1}}, \quad \forall r \in S_t^i. \tag{7}$$

Next, we choose a variable $\ell^* \notin S_t^i$ and add it to the model with the probability function

$$p_{i,\ell}^{in} = \frac{\gamma_\ell}{\sum\limits_{u \notin S_t^i} \gamma_u}, \quad \forall \ell \notin S_t^i. \tag{8}$$

Thus,

$$S_t^{i+} = \{S_t^i \setminus r^*\} \cup \{\ell^*\}$$

and we may calculate the LS solution $\hat{\beta}_t^{i+}$ for the model S_t^{i+}. 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 model or to stay at the current one. Following the scheme of simulated annealing algorithm with Metropolis–Hastings criterion, we calculate

$$q = \exp\Big(\frac{1}{nt}\big(||Y - X_{S_t^i}\hat{\beta}_t^i||_2^2 - ||Y - X_{S_t^{i+}}\hat{\beta}_t^{i+}||_2^2\big)\Big)\frac{Pr(S_t^{i+} \to S_t^i)}{Pr(S_t^i \to S_t^{i+})}$$

where

$$\begin{aligned} Pr(S_t^i \to S_t^{i+}) &= p_{i,r^*}^{out}\, p_{i,\ell^*}^{in} \\ Pr(S_t^{i+} \to S_t^i) &= p_{i^+,\ell^*}^{out}\, p_{i^+,r^*}^{in}. \end{aligned} \tag{9}$$

We are now ready for the next iteration $i+1$ by setting

$$(S_t^{i+1}, \hat{\beta}_t^{i+1}) = \begin{cases} (S_t^{i+}, \hat{\beta}_t^{i+}) & w.p \quad \min(1, q) \\ (S_t^i, \hat{\beta}_t^i) & w.p \quad \max(0, 1-q). \end{cases}$$

Along the run of the algorithm, the suggested models and their corresponding MSEs are kept. These models are used to form $\mathcal{M}(\kappa)$, and $\mathcal{G}$ can be then identified for a given value of η.

We now point out several issues for the practical application of the algorithm. First, the algorithm was described above for one single value of κ. In practice, one may run the algorithm separately for different values of κ. 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 $S_{t_1}^1$; and the vector $\gamma = (\gamma_1, ..., \gamma_p)$. Our empirical experience is that the first three can be managed without too many concerns. In particular, the algorithm should be started from more than one initial point, and the results from

each run should be kept and compared; see Sections 4 and 5. Regarding the vector γ, 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 3.2. However, in what follows we show that, under suitable conditions, the algorithm can work well even with a general choice of γ.

Define S_0, s_0 and β^0 as before and let $\mu = X\beta^0$. That is, $Y = \mu + \epsilon$. We first introduce a 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_γ for the set of positive entries in γ. 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 cardinality of A_γ and $\gamma_{min} := \min_{i \in A_\gamma} \gamma_i$ for the lowest positive entry in γ.

To motivate our next assumption, we note that, 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 $\bar{S}$ be an interesting model of size s_0–a model with not too many predictors and with a low MSE. The models we are looking for are of this nature. We facilitate the idea of $\bar{S}$ being an interesting model by assuming that $X_{\bar{S}}\hat{\beta}_{\bar{S}}$ is close to μ (in the asymptotic sense). We virtually assume that for every model of size $|\bar{S}|$, which is not $\bar{S}$, if we take out a predictor that is not part of $\bar{S}$, and replace it with a predictor from $\bar{S}$, the subspace spanned by the new model is not much further from μ, comparing with the subspace spanned by the original model. Formally, denote $\mathcal{P}_\mathcal{S}$ for the projection matrix onto the subspace spanned by the columns of the submatrix X_S.

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

$$\frac{1}{n}\left[||\mathcal{P}_{S_j^\star}\mu||_2^2 - ||\mathcal{P}_{S_{j'}^\star}\mu||_2^2\right] > 4t_0 \log c, \tag{10}$$

where $S_r^\star \equiv S \cup \{r\}$.

We note that since c could be lower than one, the right hand side of (10) can be negative. The following theorem gives conditions under which the simulated annealing algorithm is passing through an interesting model $\bar{S}$. More accurately, the theorem states that there is always strictly positive probability to pass through $\bar{S}$ in the next few moves. This result covers 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 $\bar{S}$ and other interesting models.

Let $P_t^m(S'|S)$ 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 1 *Consider the simulated annealing algorithm with $\kappa = s_0$ and with a vector γ 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 $\bar{S} \subseteq A_\gamma$ then for all $S \subseteq A_\gamma$ with $s = s_0$, for all $m \geq s_0 - |\bar{S} \cap S|$ and for large enough n,*

$$P_{t_0}^m(\bar{S}|S) > \left[\frac{c_\gamma^2}{s_0(h_\gamma - s_0)}\right]^{s_0}. \tag{11}$$

A proof is given in the appendix. Theorem 1 states that for any specification of the vector γ, such that the entries in γ are positive for all the predictors in $\bar{S}$, the probability that the algorithm would visit $\bar{S}$ 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 $\bar{S}$ in m moves.

For the classical model selection setting with $p < n$, a similar method was suggested by Brooks et al. (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. This algorithm is essentially an automatic model selection procedure.

3.2 Choosing γ

The simulated annealing algorithm described above is provided with the vector γ. The values $\gamma_1, ..., \gamma_p$ should represent the knowledge regarding the importance of the predictors, although we do not assume that any prior knowledge is available. As it can be seen in equations (7)–(8), predictors with high γ values have 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 γ includes many zeros.

One simple choice of γ is to take the absolute values of the univariate correlations of the different predictors with Y. We could also threshold the correlations in order to keep only predictors having large enough correlation (in absolute value) with Y. However, using univariate correlations is clearly problematic since it overlooks the covariance structure of the predictors in X.

Another possibility is to first use the lasso with a relatively low penalty, and then to set $\gamma_j = |\hat{\beta}_j^L|/||\hat{\beta}^L||_1$. The idea behind this suggestion is that predictors with large coefficient value may be more important for prediction of Y.

However, as discussed in Section 2, the lasso might miss some potentially good predictors. 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 γ_j using solely the elastic net. The lasso and the elastic net estimators are not model selection consistent in many situations. However, for our purpose, combining both methods together may help us get a reservoir of promising predictors.

Zou and 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 offers an additional theoretical background. 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 2 *Define X and Y as before, and define $\hat{\beta}^{EN}$ by (3). Let $X^{(1)}$ and $X^{(2)}$ be two columns of X and denote $\rho = n^{-1}(X^{(1)})^T X^{(2)}$. Assume $|\hat{\beta}_1^{EN}| \geq c_\beta$ for some $c_\beta > 0$. If $|\rho| > 1 - \lambda_2^2 c_\beta^2 / ||Y||_2^2$ then $|\hat{\beta}_2^{EN}| > 0$.*

A proof is given in the appendix. Proposition 2 gives motivation for why $\hat{\beta}^{EN}$ has typically a larger model than $\hat{\beta}^L$. It also quantifies how much correlated two predictors need to be so the elastic net would either include both predictors or none of them.

Going back to our γ vector, the next question is how to use the lasso and the elastic net in order to assign a "score" to each predictor. Let S_L and S_{EN} be the models that correspond to $\hat{\beta}^L$ 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 $\{1, ..., p\}$. Define

$$\hat{\beta}_+^L(\delta) = \arg\min_\beta \Big(\frac{1}{n} ||Y - X\beta||_2^2 + \lambda \sum_{j=1}^p \delta^{\mathbf{1}\{j \in S_+\}} |\beta_j| \Big), \qquad \delta \in [0, 1],$$

and let $S_+^L(\delta)$ be the appropriate model. In this procedure, a reduced penalty is given for predictors that $\hat{\beta}^L$ might have missed. Thus, these predictors are encouraged to enter the model, and since they 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 $\hat{\beta}_+^L(\delta)$ is a special case of $\hat{\beta}_w^L$, as defined in (2), with $w_j = \delta^{\mathbf{1}\{j \in S_+\}}$.

We demonstrate how the reduced–penalty procedure works using a toy example. A data set with $n = 30$ and $p = 50$ is simulated. We take the coefficient vector to be $\beta^0 = (0.5\ 0.5\ 1\ 1\ 1\ 0\ 0\ ...\ 0)^T$ and σ^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 $X^{(2)}$).

We suggest to measure the importance of a predictor $j \in S_+$ by the highest δ such that $j \in S_+^L(\delta)$. On the other hand, the importance of a predictor $j' \in S_L$ can be measured by the highest δ such that $j' \notin S_+^L(\delta)$ (now, smaller δ reflects j' is more important). With this in our mind, we continue to the derivation of γ.

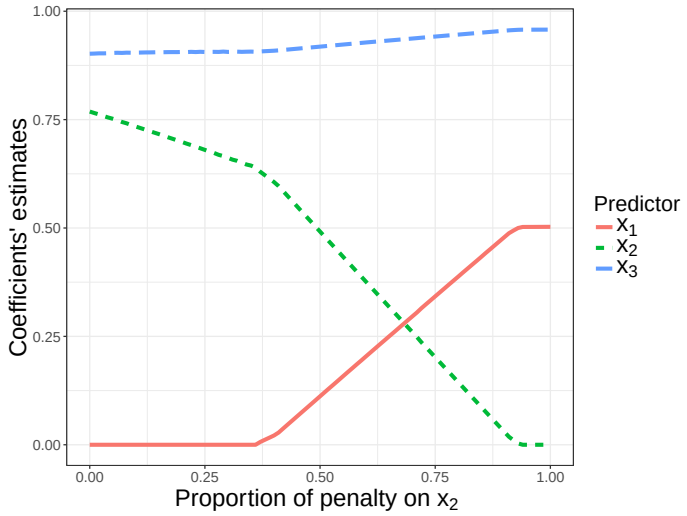


Figure 1: Toy example: coefficients' estimates for predictors $X^{(1)}, X^{(2)}$ and $X^{(3)}$ when lowering the lasso penalty for $X^{(2)}$ only. The rightmost point corresponds to a lasso procedure with equal penalties for all predictors

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

$$i^\star_j = \begin{cases} \underset{i}{\text{argmax}}\{i : \hat{\beta}^L_{+j}(\delta_i) \neq 0\} & j \notin S_L \\ \underset{i}{\text{argmax}}\{i : \hat{\beta}^L_{+j}(\delta_i) = 0\} & j \in S_L \end{cases}$$

and if the argmax is over an empty set, define $i^\star_j = 0$. We suggest to choose γ_j as follows:

$$\gamma_j = \begin{cases} 0 & j \in S_{out} \\ \delta_{i^\star_j}/2 & j \in S_+ \\ \\ 1 - \delta_{i^\star_j}/2 & j \in S_L, \end{cases}$$

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

- A predictor $j \notin S_L$ with $i^\star_j = 0$ is excluded from consideration.
- On the other hand, for a predictor $j \in S_L$, if $i^\star_j = 0$ than $\gamma_j = 1$, which is the maximal possible value. Even when the penalty for other predictors is dramatically reduced, leading to their entrance to the model, j remains part of the solution and hence it is essential for prediction of Y.
- Since predictors in S_L were picked when equal penalties were assigned to all predictors, they get priority over predictors in S_+.

- However, for two identical predictors, $X^{(j)} = X^{(j')}$ (or highly correlated predictors) such that $j \in S_L$ and $j' \notin S_L$, we get a desirable result. By Proposition 2, we know that $X'_j \in S_+$. Now, for $\delta_{h-1} < 1$ it is clear that $j' \in S_+^L(\delta_{h-1})$ and $j \notin S_+^L(\delta_{h-1})$. Therefore. $i_j^\star = i_{j'}^\star = h-1$ hence if δ_{h-1} is taken to be close to one, then $\gamma_j \simeq \gamma_{j'} \simeq 0.5$ as one might want.

Proposition 2 concerns two correlated predictors. In practice, the covariance structure of X may be much more complicated. Therefore the question 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 are also chosen? If the answer is yes (and it is, as Theorem 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 $\mathcal{G}$.

In order to reveal this property of the elastic net, we analyze $\hat{\beta}^{EN}$, the solution of (3), when assuming all the predictors in M_1 have non-zero values. Denote $M^{(-)}$ for $(M_1 \cup M_2)^c$, 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. Let $\hat{\beta}^{EN}_{M_1}$, $\hat{\beta}^{EN}_{M_2}$ and $\hat{\beta}^{EN}_{M^-}$ be the coordinates of $\hat{\beta}^{EN}$ that correspond to M_1, M_2 and $(M_1 \cup M_2)^c$, respectively. Let $\tilde{Y} = Y - \tilde{X}\hat{\beta}^{EN}_{M^-}$ be the unexplained residual 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 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 $\mathcal{P}_{M_2}$ 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*

$$\tilde{Y}^T \mathcal{P}_{M_2} X_{M_1} \hat{\beta}^{EN}_{M_1} > c_1(\lambda_1, \lambda_2, X_{M_1}, \tilde{Y}, \hat{\beta}^{EN}_{M_1}), \tag{12}$$

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 3 states that under a suitable condition, predictors belonging 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 $\mathcal{P}_{M_2} X_{M_1}\hat{\beta}^{EN}_{M_1}$ is also close to $\tilde{Y}$, would be chosen by the elastic net. Hence, these predictors are expected to have a positive value in γ, and our simulated annealing algorithm would pass through these models, provided the conditions in Theorem 1 are met. Therefore, these models are expected to appear in $\mathcal{G}$.

4. Simulation Studies

In order to examine the performance of the suggested search algorithm, we present in this section results from two simulation studies. The first investigates whether models similar in

their capability to describe the underlying population can be found by the algorithm. The second simulation study examines the out–of–sample prediction error of models included in the minimal class of models.

We present below a scenario in which, other then the model used to create the data, there are three more models, that are not nested within the underlying model, with population prediction error close to the underlying model. In practice, all of these four models are of interest.

We consider $Y = X\beta^0 + \epsilon$, $\epsilon \sim N(0, I)$ with β_j^0 equals to C for $j = 1, 2, ..., 6$ and zero for $j > 6$. C is a constant chosen to get a desired signal to noise ratio (SNR) $||X\beta^0||_2$. The predictors in X are all i.i.d. $N(0, I)$ with the exception of $X^{(7)}$ and $X^{(8)}$, which are defined by

$$X^{(7)} = \frac{2}{3}[X^{(1)} + X^{(2)}] + \xi_1, \qquad \xi_1 \sim N_n\left(0, \frac{1}{9}I\right)$$
$$X^{(8)} = \frac{2}{3}[X^{(3)} + X^{(4)}] + \xi_2, \qquad \xi_2 \sim N_n\left(0, \frac{1}{9}I\right)$$

where ξ_1 and ξ_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}. In particular, each of these models minimizes the population prediction error $E||Y - X_S\beta_S||_2^2\}$ for models with its size. That is, Model (I) minimizes the population prediction error for models of size $k = 6$, Model (II) minimizes the prediction error for models of size $k = 4$, and both models (III) and (IV) minimize the prediction error for models of size $k = 5$. Furthermore, it can be shown, for example, that the population prediction error of Model (III) is just 3.7% larger than the error of true model for $SNR = 1$, 14.8% for $SNR = 2$, and larger values for $SNR \geq 4$

Note that while models (II)–(IV) do not describe the data as well as Model (I), they contain less predictors. Furthermore, in an hypothetical real-life situation where future measurements were to be made on the predictors and the outcome, and, for the sake of the example, $X^{(1)}$ or $X^{(2)}$ were more expensive to measure, models (II)–(IV) could have provided a frugal alternative, while preserving a reasonable prediction error.

As part of this simulation study, for each simulated data set, we do the following:

1. Obtain γ as explained in Section 3.2. The tuning parameter of the lasso is taken to be the minimizer of the cross–validation MSE. For the elastic net, α in (4) is taken to be 0.4.

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

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

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

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.

A 1000 simulated data sets were generated for each different scenario: 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 might expect. For large SNR, the models are chosen more frequently. However, models (III) and (IV) are competing, in the sense that they both include five predictors. Even for large SNR, each of the models, (III) and (IV), is chosen in about 50% of the cases. Therefore, as recommended in Section 3.1, we repeat the simulations while starting the algorithm from different initial points, that is, different initial models.

Figure 2 presents comparison between running the algorithm from one and three starting points. When we initiated the algorithm from three different models, the results improved for all models, and in particular for models (III) and (IV). The results described in this section are quite similar to the results we obtained when forming $\mathcal{G}(\kappa, \eta)$ as defined in (6), for each $\kappa = 4, 5, 6$ separately and using arbitrary small values of η.

We now turn to describe the second simulation study, that concerns out–of–sample prediction of the models found by the algorithm. Here we consider three scenarios. The true model consisting 20 predictors for all scenarios. In Scenario 1, X consists p iid normally distributed predictors with mean zero and variance one. The 20 first entries in β^0 are created from iid normal distribution, and then rescaled to achieve the desired SNR. In Scenario 2,

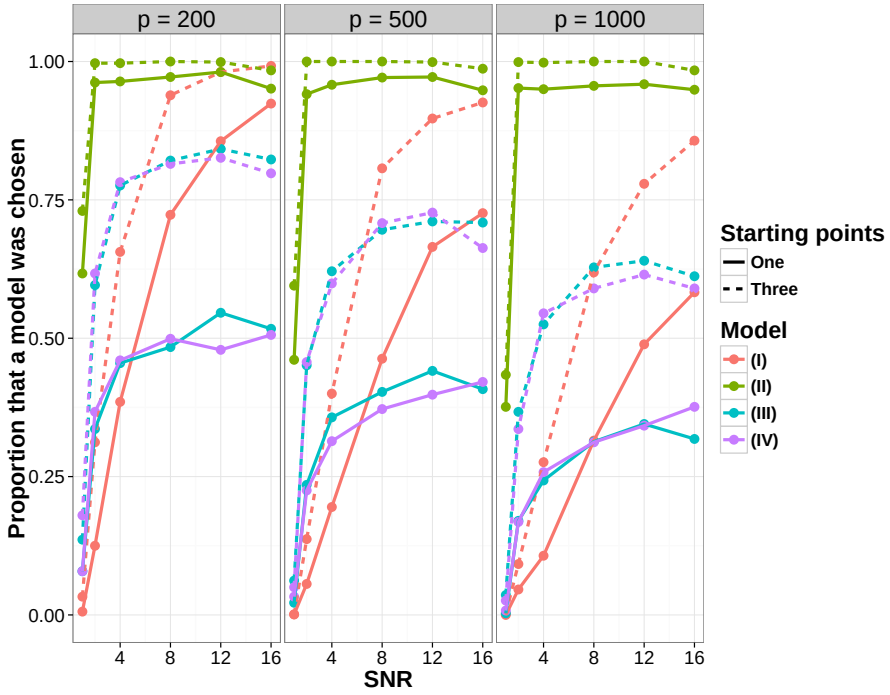


Figure 2: Proportion that each model is chosen as one of top five models for different number of potential predictors (p) and various SNR values. There is an apparent improvement when running the algorithm from three starting points.

β^0 is created in the same as in Scenario 1, but a more complex correlation structure is considered for X. First, define $\Sigma^{(m)}$ to be an $m \times m$ covariance matrix defined so for all $i = 1, ..., m$ and $j = 1, ., ., m$, $\Sigma_{ij}^{(m)} = 0.75^{|i-j|}$. The subvector containing the first 5 predictors, $(X^{(1)}, X^{(2)}, ..., X^{(5)})$ is created with the covariance matrix $\Sigma^{(5)}$. Then, each of the next 10 predictors $X^{(j)}, j = 6, 7..., 15$ is correlated with two predictors that are not part of the true model. For example, $(X^{(6)}, X^{(21)}, X^{(22)})$ is simulated from $N_3(0, \Sigma^{(3)})$ and $(X^{(7)}, X^{(23)}, X^{(24)})$ is simulated from $N_3(0, \Sigma^{(3)})$. The remaining predictors are simulated as iid $N(0, 1)$. In Scenario 3, X is created as in Scenario 2, but the coefficients in the true model are now all equal. That is, $\beta_j^0 = C, j = 1, 2, ..., 20$, with C being a constant chosen to get a desired SNR.

As in the previous study, $Var(\epsilon) = I, n = 100$ and we consider $p = 200, 500, 1000$. We report the results for SNR=1,2,4,6,8. For each synthetic data set, the simulated annealing is used for $\kappa = 15, 20, 25$. The algorithm is initiated three times per κ value, and the best five models per run are kept, and then combined to have a minimal class of models containing up to 15 models (per κ value). Other tuning parameters are chosen as in the first simulation study. Figure 3 compares mean (over the synthetic data sets) out–of–sample prediction error for the best model (in terms of prediction error) within the minimal class of models, the average of out–of–sample prediction error among the models in the minimal class, as well as the errors of the lasso, elastic net and adaptive lasso. Results are presented

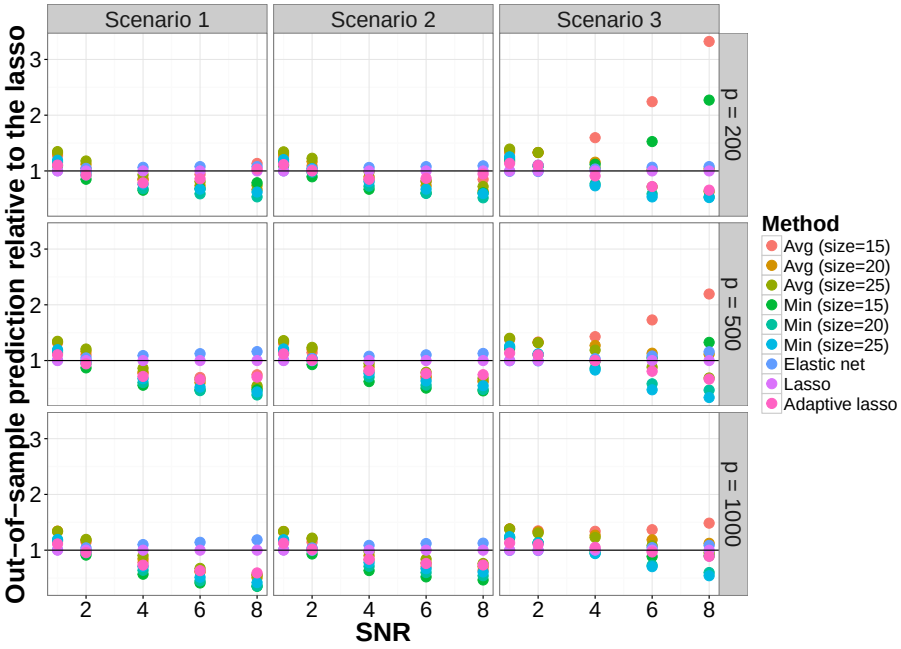
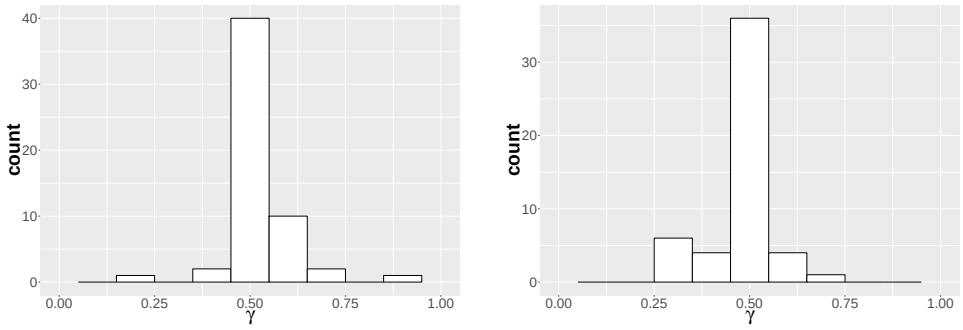


Figure 3: Out–of–sample prediction error for different scenarios, different number of potential predictors (p) and various SNR values. Methods compared are lasso, elastic net and adaptive lasso, as well as the average (avg) and minimal (min) error in the minimal class of models, with model size of 15,20 and 25 predictors. The classes were built by combining 5 best models obtained from three runs of the simulated annealing search algorithm.

relative to the out–of–sample prediction error of the lasso. For SNR=1, the lasso performs the best, for all three scenarios considered, although the other methods show comparable performance. Under Scenarios 1 and 2, when the non–zero parameters are created from a normal distribution, so some predictors are stronger and others are weaker, the best model identified by the minimal class, for either $\kappa = 20$ or $\kappa = 25$, outperforms other methods for SNR>1. Under Scenario 3, where all coefficients (in the true model) are equal, the same pattern is observed for SNR>4. For large SNR values, under Scenario 3, the models in the minimal class of size 15 are underperforming. While under Scenarios 1 and 2, the predictors missed by the minimal class with $\kappa = 15$ are likely to be those with small effect, under Scenario 3 all predictors have an equal effect. We also considered scenarios with X simulated from $N_p(0, \Sigma^{(p)})$. The results are similar to those observed for Scenario 1 and hence are omitted.

5. Real data sets

We demonstrate the utility of using a minimal class of models in the analysis of two real data sets. The tuning parameters of the lasso and the elastic net were taken to be the same as in Section 4. The tuning parameters of the simulated annealing algorithm were



(a) Riboflavin data (b) Air pollution data

Figure 4: Histograms of the values of γ for positive entries only in the two data set analysis examples.

$T = 10 \times (0.7^1, 0.7^2, ..., 0.7^{20})$, $\Delta = (0, 0.01, 0.02, ..., 0.98, 0.99, 1)$, and $N_t = N = 100$ for all $t \in T$.

5.1 Riboflavin

We use a high–dimensional data about the production of riboflavin (vitamin B2) in Bacillus subtilis that were recently published (Bühlmann et al., 2014). The data consist $p = 4088$ predictors. These are measures of log expression levels of genes in $n = 71$ observations. The target variable is the (log) riboflavin production rate.

The lasso model S_L includes 40 predictors (and intercept), and the elastic net model S_{EN} includes 59 predictors. In total, we get 61 different predictors (i.e., genes). Panel (a) of Figure 4 presents the histogram of the positive values in γ.

We run the algorithm from three random starting points for each model size between 1 and 10. We keep the five best models for each size and starting point, 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. Following a referee comment, we note here that in practice models with low model size may have unacceptably low R^2, and this can be checked in practice. In our analysis here we keep these models since the R^2 values are all larger than 0.35 and to ease the exposition, so we would be able to compare the search results to models obtained by other methods. The models found by other methods, for the riboflavin data, are of a relatively small size; see Bühlmann et al. (2014)

The following insights are drawn from examining more carefully the models we obtained (see Table 1 in the online 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 our models. We can infer that while

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

Table 2: Riboflavin data: Number of unique models for each model size after running the algorithm from 3 different starting points

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 all 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 these genes only, 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 (Yuan and Lin, 2006).

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

- The importance of genes number 792,1131, and possibly others, should be also examined since each of them appears in a variety of different models.

We now compare our results to 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 models. If one constructs his model using the stability selection (Meinshausen and Bühlmann, 2010) as a screening process for the predictors, he would get a model consisting 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 MSE for a model in our 8 models of size 3 is 0.2047 while the MSE of the model suggested using the stability selection is 0.2703, more than 30% difference!

5.2 Air pollution

We now demonstrate how the proposed procedure can be used for traditional, purportedly simpler, problem. The air pollution data set (McDonald and Schwing, 1973) includes 58 Standard Metropolitan Statistical Areas (SMSAs) of the US (after removal of outliers). The outcome variable is age-adjusted mortality rate. There are 15 potential predictors including air pollution, environmental, demographic and socioeconomic predictors. Description of the predictors is given in Table 4 in the appendix.

There is no guarantee that the relationship between the predictors and the outcome variable has a 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.

High–dimensional regression model that includes transformations and interactions has been dealt with in the literature. For example, by using two step procedures (Bickel et al., 2010) or by solving a relevant optimization problem (Bien et al., 2013). Our procedure has a different goal, since we are not looking for the best predictive model, but rather for meaningful insights about the data.

Following the lasso and elastic net step, we are left with 51 predictors with positive γ_j (6 untransformed predictors, 4 log transformations, 6 square root transformations, 9 power of two transformations and the rest are interactions). Panel (b) of Figure 4 presents the histogram of the positive values in γ.

We modify the search algorithm to make it produce results typical to an analysis involving interaction terms. In particular, we search for models such that an interaction term is included only if at least one of the corresponding main terms is included in the model. In order to do so, the definitions in (7) and (8) are modified such that the probability of proposing an interaction term is zero if none of the corresponding main effects is in this model (excluding the variable chosen to be taken out). On the other hand, main effects are "protected" of being suggested to be excluded from the model, if any interaction involving them is part of the current model. We achieve this by setting to zero the probability of suggesting a predictor when an interaction term involving this predictor is included.

Following our earlier comment on considering satisfactory models, and since we are looking for more complex models in this example, considering the option of transformations and interactions, we run the algorithm for $\kappa = 5, 6, ..., 15$, for each κ, from three starting points, and then we keep the 5 best models. In total, we get 164 unique models. Table 3 summarizes the results for prominent main effect predictors, that is, predictors that appear in at least third of the models we obtained. The table presents a matrix of the joint frequency of each two predictors. Each cell in the table is the number of models including both the predictor listed in the row and the predictor listed in the column. The diagonal is simply the number of models that a predictor appears in. Considering Table 3, 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. The percentage of non–white population also appears in most models. In almost half of them, it appears untransformed, but the same could be said about this predictor after square root transformation. However, in less than 10% of the models, this predictor appears in both forms. We conclude that this predictor should be used for prediction of the mortality rate, but the question of transformation remains unsolved. Similar comments can be made about hydrocarbon pollution.

We turn to the interactions. Because of the way we searched for interactions, it becomes "harder" for an interaction term to enter a model, and we therefore analyze them separately. Out of the 26 interactions considered, none clearly stood out above the rest. Five of the interaction terms appear in about 15% of the models. Two of those involve the percentage of non–white population, one with the prevalence of low–income, and the other with percentage of housing units with all facilities. Another of those five interaction terms is the interaction between percentage of elderly population and the average temperature in January. The latter main effect appears in about 50 models, and hence did not make the cutoff for Table 3. The percentage of elderly population was not included in either the lasso or elastic net

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) **prec**	57	25	25	52	27	29	18	18	27
(2) **nwht**		81	33	75	34	15	39	37	39
(3) log(**HC**)			61	60	21	34	9	29	33
(4) log(**NOx**)				158	67	85	64	56	80
(5) $\sqrt{\textbf{jant}}$					71	37	25	23	37
(6) $\sqrt{\textbf{nwht}}$						85	33	27	45
(7) $\sqrt{\textbf{HC}}$							66	29	22
(8) $\textbf{jult}^2$								61	24
(9) $\textbf{educ}^2$									83

Table 3: Frequency that each two predictors together in the 164 models. The diagonal is simply the number of models that a predictor appears in. For example, in 67 models both log(**NOx**) and $\sqrt{\textbf{jant}}$ appear.

preliminary steps. However, the absence of age related effect is not so surprising since the outcome variable, the mortality rate, is age corrected. Nevertheless, since the interaction term appeared in about half of the models that included January temperature as a predictor, it can be considered if the main effect is also considered.

6. Discussion

Model selection consistency is an ambitious goal to achieve when dealing with high–dimensional data. A "minimal class of 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 a simulated annealing method. Under suitable conditions, that are outlined in Theorem 1, the algorithm passes through models of interest.

A score for each predictor is given using the lasso, the elastic net and a reduced–penalty lasso. These scores are used by the search algorithm. They are not necessarily optimal 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. A simulation study demonstrated the capability of the search algorithm to detect relevant models.

One possible limitation of the proposed algorithm is the number of tuning parameters to be chosen. In our simulation studies and data analyses, we have found the algorithm to be quite robust to the parameters' specification. For ease of applications, we now list suggested default values for some of the parameters. The tuning parameter of the lasso λ, can be chosen by cross validation. Since the subsequent use of the elastic net is to bring into surface potential predictors, relatively low value for α in (4) can be taken, say $\alpha \in (0.2, 0.5)$. A possible choice for Δ is $(0, 0.01, 0.02, ..., 1)$. As always when using simulated annealing, T and N_t may be more complicated to choose. We used for all simulations and data analyses the sequence $T = (10 \times 0.7^1, 10 \times 0.7^2, ..., 10 \times 0.7^{20})$ and $N_t = 100$. This leaves us with the decision of which models sizes κ to consider. A preliminary run with, say $\kappa = 1, 2, ..., 10$

and, for example, comparing the MSEs for $\kappa = 5$ and $\kappa = 10$, may imply if there is a substantial benefit when adding variables, before moving to larger models. Similarly to many algorithms, the proposed simulated annealing algorithm should be initiated from multiple points.

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 $p > n$ regime. Therefore, we suggest to abandon the search for this "holy grail", and to analyze the class of minimal models instead. While retreating from the choice of a single model, and instead analyzing group of models, offers advantages, it may complicate standard data analysis in at least two ways, even when putting computational issues aside. The first is a practical one and concerns the question of which analyses should be carried out once a minimal class of models, or at least its sample version, is obtained. Model averaging can be used to estimate association parameters. Similarly, predictions from multiple models can be averaged. More precise predictions or estimates may be obtained by coupling the obtained models with weights, possibly using the empirical MSEs. More robust aggregation can be obtained by using medians instead of averages, to deal with the fact that each model consists of a different number of predictors. Another, less formal, data analysis strategy is to combine subject matter knowledge with wealth of many models. We have demonstrated such exploratory analyses in Section 5, by comparing the proportion of models each predictor, or multiple predictors, are included.

The second issue is a conceptual one. If we indeed quit from searching for a single model, i.e., a single group of predictors, what are our assumptions about the underlying mechanism that the data came from. Do we believe such a mechanism exists? One answer is that we do not need to think about how the data was created, but how can we describe the data. This, however, offers only a partial answer because we are focused on exploratory analysis, while model selection is of interest from inferential perspective as well. If more than one model describes the data adequately, which one is the true one? Does it matter? While most data are not created following one's computer program, disbelieving in the existence of a true model certainly gives rise to further questions.

It is well known that achieving good prediction and successful model selection simultaneously, in a reasonable computation time, is impossible, especially in the high–dimensional setting. We therefore suggested here to make a compromise. Our approach is not necessarily optimal for prediction, nor for model selection. However, it offers a data analysis method that takes into account the uncertainty in model selection, but ensures reasonable prediction accuracy. This method can be used for either prediction, parameter estimation or model selection.

Acknowledgments

We thank two anonymous reviewers for useful comments and suggestions that improved the paper. This research was supported by ISF grant 1770/15

Appendix

Appendix A. Proofs

A.1 Proof of Theorem 1

We start with the following lemma.

Lemma 4 *Assume $Y = \mu + \epsilon$ and assume also (A1)–(A4). Let $\mathcal{S}_k = \{S : |S| = k, \hat{\beta}_S = (X_S^T X_S)^{-1} X_S^T Y\}$ be the set of all models with k variables, such that $\hat{\beta}_S$, the LS estimate, is unique. Denote $S_j^\star = S \cup \{j\}, j \notin S$ for a model that includes S and additional variable j not in S. We have*

$$\max_{\substack{S \in \mathcal{S}_k \\ 1 \le j \le p}} \epsilon^T (X_{S_j^\star} \hat{\beta}_{S_j^\star} - X_S \hat{\beta}_S) = o_p(n)$$

Let ξ_j be the vector of coefficients obtained by regressing $X^{(j)}$, the j^{th} column in X, on X_S and let $\mathcal{P}_j$ 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,

$$\mathcal{P}_j = \frac{(X^{(j)} - X_S \xi_j)(X^{(j)} - X_S \xi_j)^T}{||X^{(j)} - X_S \xi_j||_2^2}.$$

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

$$\begin{aligned} X_{S_j^\star} \hat{\beta}_{S_j^\star} &= X^{(j)} \hat{\beta}^j_{S_j^\star} + X_S \hat{\beta}^{-j}_{S_j^\star} \\ &= (X^{(j)} - X_S \xi_j) \hat{\beta}^j_{S_j^\star} + X_S (\hat{\beta}^{-j}_{S_j^\star} + \xi_j \hat{\beta}^j_{S_j^\star}) \\ &= (X^{(j)} - X_S \xi_j) \hat{\beta}^j_{S_j^\star} + X_S \hat{\beta}_S \\ &= \mathcal{P}_j y + X_S \hat{\beta}_S. \end{aligned}$$

Therefore,

$$\epsilon^T (X_{S_j^\star} \hat{\beta}_{S_j^\star} - X_S \hat{\beta}_S) = \epsilon^T \mathcal{P}_j \mu + \epsilon^T \mathcal{P}_j \epsilon.$$

Now, since $||\mathcal{P}_j \mu||_2^2 \le ||\mu||_2^2 = O(n)$, we get that for all j, $\epsilon^T \mathcal{P}_j \mu = O_p(\sqrt{n})$. Next, let $Z_1, ..., Z_{p^{k+1}}$ be $N(0, \sigma^2)$ random variables and observe that the approximate size of the set $\{S_k\} \times \{1, ..., p\}$ is p^{k+1}. We have for any $a > 0$

$$P\left(\max_{\substack{S \in \mathcal{S}_k \\ 1 \le j \le p}} \frac{1}{n} \epsilon^T \mathcal{P}_j \epsilon \ge a \right) \le P\left(\max_{1 \le j \le p^{k+1}} |Z_j| \ge \sqrt{\frac{an}{\sigma^2}} \right) \le \sigma \sqrt{\frac{2(k+1)\log p + o(1)}{an}}.$$

Now, since $p = n^\alpha$ and $k = o(n/\log n)$ we get that

$$P\left(\max_{\substack{S \in \mathcal{S}_k \\ 1 \le j \le p}} \frac{1}{n} \epsilon^T \mathcal{P}_j \epsilon \ge a \right) = o(1)$$

and we are done. □

We can now move to the proof of Theorem 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 $g \in S^c \cap \bar{S}$. Denote $S' = \{S/\{j\}\} \cup \{g\}$ for this new model. We have

$$\begin{aligned} &Q_t(S, g, j) \\ &\quad = Pr(S \to S') \min \left[1, \exp \left(\frac{||Y - X_S \hat{\beta}_S||_2^2 - ||Y - X_{S'} \hat{\beta}_{S'}||_2^2}{t} \right) \frac{Pr(S' \to S)}{Pr(S \to S')} \right] \end{aligned} \tag{13}$$

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

$$\begin{aligned} c_\gamma(h_\gamma - s_0) \leq \sum_{u \notin S} \gamma_u \leq h_\gamma - s_0 \\ s_0 \leq \sum_{v \in S} \frac{1}{\gamma_v} \leq \frac{s_0}{c_\gamma}. \end{aligned} \tag{14}$$

Now, by substituting (14) into (7)–(9) we get

$$\begin{aligned} Pr(S \to S') &= \frac{\gamma_g}{\sum_{u \notin S} \gamma_u} \frac{1/\gamma_j}{\sum_{v \in S} \frac{1}{\gamma_v}} \geq \frac{c_\gamma^2}{s_0(h_\gamma - s_0)}, \\ \frac{Pr(S' \to S)}{Pr(S \to S')} &= \frac{\gamma_j^2}{\gamma_g^2} \frac{\sum_{u \notin S} \gamma_u \sum_{v \in S} \frac{1}{\gamma_v}}{\sum_{u \notin S'} \gamma_u \sum_{v \in S'} \frac{1}{\gamma_v}} \geq c_\gamma^4. \end{aligned} \tag{15}$$

Next, we have

$$\begin{aligned} &\frac{1}{n}||Y - X_S \hat{\beta}_S||_2^2 - \frac{1}{n}||Y - X_{S'} \hat{\beta}_{S'}||_2^2 \\ &\quad = \frac{1}{n} \left[(Y - X_{S'} \hat{\beta}_{S'}) + (Y - X_S \hat{\beta}_S) \right]^T \left(X_{S'} \hat{\beta}_{S'} - X_S \hat{\beta}_S \right) \\ &\quad = \frac{1}{n} Y^T \left(X_{S'} \hat{\beta}_{S'} - X_S \hat{\beta}_S \right) \\ &\quad = \frac{1}{n} \mu^T \left(X_{S'} \hat{\beta}_{S'} - X_S \hat{\beta}_S \right) + \frac{1}{n} \epsilon^T \left(X_{S'} \hat{\beta}_{S'} - X_S \hat{\beta}_S \right) \\ &\quad = \frac{1}{n} \mu^T \left(X_{S'} \hat{\beta}_{S'} - X_S \hat{\beta}_S \right) + \Delta_n(S, S') \end{aligned} \tag{16}$$

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 prediction using this estimator with Y is larger. Now, denote $S'' = S' \cup S$. We have

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

and if we apply Lemma 4 twice we get that $\Delta_n(S,S') = o_p(1)$. Now, regarding the first term in (16),

$$\begin{aligned}\frac{1}{n}\mu^T\left(X_{S'}\hat{\beta}_{S'} - X_S\hat{\beta}_S\right) &= \frac{1}{n}\mu^T\left(\mathcal{P}_{\mathcal{S}'}y - \mathcal{P}_{\mathcal{S}}y\right)\\ &= \frac{1}{n}\left(||\mathcal{P}_{\mathcal{S}'}\mu||_2^2 - ||\mathcal{P}_{\mathcal{S}}\mu||_2^2\right) + \Delta_n'(S,S')\end{aligned} \tag{17}$$

where $\Delta_n'(S,S') = \frac{1}{n}\mu^T\left[\mathcal{P}_{\mathcal{S}'}\epsilon - \mathcal{P}_{\mathcal{S}}\epsilon\right]$. The content of the proof of Lemma 4 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

$$\frac{1}{n}\left(||Y - X_S\hat{\beta}_S||_2^2 - ||Y - X_{S'}\hat{\beta}_{S'}||_2^2\right) \geq 4t\log c_\gamma. \tag{18}$$

Now, by substituting (15) and (18) into (13) we get that for large enough n,

$$Q_{t_0}(S,g,j) \geq \frac{c_\gamma^2}{s_0(h_\gamma - s_0)}$$

for all $S \neq \bar{S}$, $j \in S \cap \bar{S}^c$ and $g \in S^c \cap \bar{S}$. (11) follows from this immediately since for any integer m and for all $S \neq \bar{S}$,

$$P_{t_0}^m(S'|S) \geq \min_{\substack{\{S:S\cap\bar{S}=\emptyset\}\\ j\in S\cap\bar{S}^c\\ g\in S^c\cap\bar{S}}}[Q_t(S,g,j)]^{s_0} \geq \left[\frac{c_\gamma^2}{s_0(h_\gamma - s_0)}\right]^{s_0}.$$

A.2 Proof of Proposition 2

Recall that the elastic net estimator $\hat{\beta}^{EN}$ minimizes

$$||Y - X\beta||_2^2 + \lambda_1|\beta| + \lambda_2||\beta||_2^2 \tag{19}$$

Now, WLOG assume that $\hat{\beta}^{EN}$ is a solution such that $\hat{\beta}_1^{EN} > 0$. For convenience, we omit the "EN" superscript from now on (i.e., $\hat{\beta} = \hat{\beta}^{EN}$). Define the subspace

$$\mathcal{B} := \{\beta : \forall i \neq 1,2 \;\; \beta_i = \hat{\beta}_i, \;\; \beta_1 = \tau\hat{\beta}_1, \;\; \beta_2 = (1-\tau)\hat{\beta}_1\}. \tag{20}$$

If the minimum of (19) over $\mathcal{B}$ is obtained for $\tau \neq 1$, then given that $X^{(1)}$ is part of the elastic net model, predictor $X^{(2)}$ is also part of this model.

WLOG, write down X as $X = (X_{(12)} \quad X_{-(12)})$ where $X_{(12)} = (X^{(1)} \quad X^{(2)})$ are the first two columns of X and $X_{-(12)}$ are the rest of its columns. Similarly, we have $\beta^T = (\beta_{(12)}^T \quad \beta_{-(12)}^T)$ where $\beta_{(12)}$ is the first two entries in the vector β and $\beta_{-(12)}$ is the rest of the vector. Define $\tilde{Y} = Y - X_{-(12)}\beta_{-(12)}$. We can rewrite (19) as

$$||\tilde{Y} - X_{(12)}\beta_{(12)}||_2^2 + \lambda_1(|\beta_{-(12)}| + |\beta_{(12)}|) + \lambda_2(||\beta_{-(12)}||_2^2 + ||\beta_{(12)}||_2^2) \tag{21}$$

If the minimum of (21), on $\mathcal{B}$, is achieved at $0 < \tau^* < 1$ then $\hat{\beta}_2$ must be non zero. Minimizing (21) on $\mathcal{B}$ is essentially minimizing

$$-2\tilde{Y}^T X_{(12)}\beta_{(12)} + ||X_{(12)}\beta_{(12)}||_2^2 + \lambda_2||\beta_{(12)}||_2^2 \tag{22}$$

on $\mathcal{B}$. Now, by the definition of $\mathcal{B}$ in (20) and using simple algebra we get that (22) equals to

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

This is a quadratic function of τ, and by equating its derivative to zero we get that

$$\tau^*=\frac{1}{2}-\frac{\tilde{Y}^T(X^{(2)}-X^{(1)})}{2\hat{\beta}_1(\lambda_2+1-\rho)}$$

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

$$\left|\frac{\tilde{Y}^T(X^{(2)}-X^{(1)})}{\hat{\beta}_1(\lambda_2+1-\rho)}\right|<1. \tag{23}$$

Since $||X^{(2)}-X^{(1)}||_2^2=2(1-\rho)$ we have

$$|\tilde{Y}^T(X^{(2)}-X^{(1)})|\leq\sum_{i=1}^{n}|\tilde{Y}_i||X_i^{(2)}-X_i^{(1)}|\leq||\tilde{Y}||_2\sqrt{2(1-\rho)},$$

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 $||\tilde{Y}||_2\leq||Y||_2$. 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 ρ's. For $\lambda_2<\frac{u^2}{2}$ we have

$$\sqrt{1-\rho}>\frac{1}{\sqrt{2}}(u+\sqrt{u^2-2\lambda_2}), \tag{24}$$

$$\sqrt{1-\rho}<\frac{1}{\sqrt{2}}(u-\sqrt{u^2-2\lambda_2}). \tag{25}$$

The RHS of (24) is larger than 1 if $\lambda_2<\sqrt{2}u-1$. That is, there is no suitable ρ 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 ρ,

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

or alternatively,

$$\rho>1-\frac{u^2}{2}\left(1-\sqrt{1-\frac{2\lambda_2}{u}}\right)^2$$

and by Taylor expansion for $2\lambda_2/u$ we get

$$\rho > 1 - \frac{\lambda_2^2}{2u^2}$$

■

A.3 Proof of Theorem 3

The proof is similar to the proof of Proposition 2. Let $\hat{\beta} = \hat{\beta}^{EN}$ be the elastic net estimator and denote $\hat{\beta}_M$ for the values in $\hat{\beta}$ corresponding to the set of predictors M. We can partition the set of potential predictors $\{1, 2, ..., p\}$ to four disjoint subsets: $M^{(-)}$; $M_1 \cap M_2^c$; $M_1^c \cap M_2$ and $M_1 \cap M_2$. We replace (20) with

$$\mathcal{B} := \{\beta : \beta_{M^{(-)}} = \hat{\beta}_{M^{(-)}} \;\; \beta_{M_1 \cap M_2} = \hat{\beta}_{M_1 \cap M_2}, \tag{26}$$

$$\beta_{M_1 \cap M_2^c} = \tau \hat{\beta}_{M_1 \cap M_2^c}, \;\; \beta_{M_1^c \cap M_2} = (1-\tau)\Theta' \hat{\beta}_{M_1 \cap M_2^c}\}. \tag{27}$$

where β_M is defined as the values in $\hat{\beta}$ corresponding to the set M and Θ' is the matrix of coefficients obtained from regressing $X_{M_1 \cap M_2^c}$ on $X_{M_1^c \cap M_2}$. We define Θ to be an augmented version of Θ', which we obtain by regressing X_{M_1} on X_{M_2}. That is,

$$X_{M_2}\Theta = \mathcal{P}_{M_2} X_{M_1} \tag{28}$$

Note that on $\mathcal{B}$,

$$X\beta = \tilde{X}\hat{\beta}_{M^{(-)}} + \tau X_{M_1}\hat{\beta}_{M_1} + (1-\tau)X_{M_2}\Theta\hat{\beta}_{M_1}$$

Recalling that $\tilde{Y} = Y - \tilde{X}\hat{\beta}_{M^{(-)}}$, minimizing (3) on $\mathcal{B}$ is equivalent to minimize

$$\begin{aligned} ||\tilde{Y} - \tau X_{M_1}\hat{\beta}_{M_1} - (1-\tau)X_{M_2}\Theta\hat{\beta}_{M_1}||_2^2 + \lambda_1[\tau||\hat{\beta}_{M_1}||_1 + (1-\tau)||\Theta\hat{\beta}_{M_1}||_1] \\ + \lambda_2[\tau^2||\hat{\beta}_{M_1}||_2^2 + (1-\tau)^2||\Theta\hat{\beta}_{M_1}||_2^2] \end{aligned} \tag{29}$$

as a function of τ. Using a first–order condition and substituting (28) we find that (29) is minimized for

$$\tau^* = \frac{-(\tilde{Y} - \mathcal{P}_{M_2}X_{M_1}\hat{\beta}_{M_1})^T(I - \mathcal{P}_{M_2})X_{M_1}\hat{\beta}_{M_1} + \frac{\lambda_1}{2}(||\hat{\beta}_{M_1}||_1 - ||\Theta\hat{\beta}_{M_1}||_1) - \lambda_2||\Theta\hat{\beta}_{M_1}||_2^2}{||(I - \mathcal{P}_{M_2})X_{M_1}\hat{\beta}_{M_1}||_2^2 - \lambda_2||\hat{\beta}_{M_1}||_2^2 - \lambda_2||\Theta\hat{\beta}_{M_1}||_2^2}. \tag{30}$$

Before we continue, note that if $X_2 = X_1$ then Θ is the identity matrix and $\mathcal{P}_{M_2}X_{M_1} = X_1$. Substituting these facts into (30), we get that $\tau^* = \frac{1}{2}$ as one might expect. 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

$$\begin{aligned} -\tilde{Y}^T(I - \mathcal{P}_{M_2})X_{M_1}\hat{\beta}_{M_1} + ||(I - \mathcal{P}_{M_2})X_{M_1}\hat{\beta}_{M_1}||_2^2 - \frac{\lambda_1}{2}||\Theta\hat{\beta}_{M_1}||_1 \\ > -\frac{\lambda_1}{2}||\hat{\beta}_{M_1}||_1 - \lambda_2||\hat{\beta}_{M_1}||_2^2 \end{aligned}$$

which is certainly true if

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

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

Appendix B. Supplementary table for Section 5.2

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
h **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 $<$ 3,000 dollars in urbanized areas
HC	Relative pollution potential of hydrocarbon
NOx	Relative pollution potential of nitric oxides
SUL	Relative pollution potential of sulfur dioxide
hum	Annual average percentage of relative humidity at 1pm

Table 4: Potential predictors for mortality rate in Section 5.2

References

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Peter J Bickel and Mu Cai. Discussion of Sara van de Geer: Generic chaining and the ℓ_1 penalty. *Journal of Statistical Planning and Inference*, 143(6):1013–1018, 2012.

Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*, volume 6, pages 56–69. Institute of Mathematical Statistics, 2010.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111–1141, 2013.

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

S. P. Brooks and B. J. T. Morgan. Optimization using simulated annealing. *The Statistician*, pages 241–257, 1995.

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

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1: 255–278, 2014.

Florentina Bunea, AlexandreB. Tsybakov, and MartenH. Wegkamp. Aggregation and Sparsity Via $\ell 1$ Penalized Least Squares. In *Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 379–391. Springer Berlin Heidelberg, 2006. URL `http://dx.doi.org/10.1007/11776420_29`.

Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22, 2010.

Eitan Greenshtein and Ya'Acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

Nils Lid Hjort and Gerda Claeskens. Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899, 2003.

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

Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, pages 382–401, 1999.

Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603, 2008.

Jinzhu Jia and Bin Yu. On model selection consistency of the Elastic Net when $p >> n$. *Statistica Sinica*, 20:595–611, 2010.

Scott Kirkpatrick, D. Jr. Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

Gary C McDonald and Richard C Schwing. Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15(3):463–481, 1973.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

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

Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270, 2009.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488), 2009.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

Robert B O'Hara and Mikko J Sillanpää. A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.

Philippe Rigollet and Alexandre B Tsybakov. Sparse estimation by exponential weighting. *Statistical Science*, 27(4):558–575, 2012.

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

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 71:879–898, 2012.

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

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

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

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

Cun-Hui Zhang and Jian Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

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

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

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

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