Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Choosing the Best Subset Is Not the Best Choice

Moritz Hanke (hanke@leibniz-bips.de)
Leibniz Institute for Prevention Research and Epidemiology - BIPS
Louis Dijkstra
Leibniz Institute for Prevention Research and Epidemiology - BIPS
Ronja Foraita
Leibniz Institute for Prevention Research and Epidemiology - BIPS
Vanessa Didelez
Leibniz Institute for Prevention Research and Epidemiology - BIPS

Research Article

Keywords: variable selection, high dimensional, best subset selection

Posted Date: September 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-743866/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

RESEARCH

When choosing the best subset is not the best choice

Moritz Hanke*†, Louis Dijkstra†, Ronja Foraita and Vanessa Didelez

*Correspondence:
hanke@leibniz-bips.de
Department Biometry and Data Management, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Achterstraße 30, 28359 Bremen, Germany
Full list of author information is available at the end of the article
†Shared first authorship

Abstract

Background: Variable selection in linear regression settings is a much discussed problem. Best subset selection (BSS) is often considered as an intuitively appealing 'gold standard', with its use being restricted mainly by its NP-hard nature. Instead, alternatives such as the least absolute shrinkage and selection operator (Lasso) or the elastic net (Enet) have become methods of choice in high-dimensional settings. A recent proposal represents BSS as a mixed integer optimization problem so that much larger problems have become feasible in reasonable computation time. This has been exploited to study the prediction performance of BSS and its competitors. Here, we present an extensive simulation study assessing, instead, the *variable selection* performance of BSS compared to forward stepwise selection (FSS), Lasso and Enet. The analysis considers a wide range of settings that are challenging with regard to dimensionality, signal-to-noise ratio and correlations between relevant and irrelevant direct predictors. As measure of performance we used the best possible F1 score for each method so as to ensure a fair comparison irrespective of any criterion for choosing the tuning parameters .

Results: Somewhat surprisingly, it was *only* in settings where the signal-to-noise ratio was high and the variables were (nearly) uncorrelated that BSS reliably outperformed the other methods. This was the case even in low dimensional settings where the number of observations exceeded the number of variables by a factor of ten. Further, the FSS approach performed nearly identically to BSS.

Conclusion: Our results shed a new light on the usual presumption of BSS being, in principle, the best choice for variable selection. More attention needs to be payed to the data generating process when considering variable selection methods. Especially for correlated variables, convex alternatives like Enet are not only faster but also appear to be more accurate in practical settings.

Keywords: variable selection; high dimensional; best subset selection

Introduction

Selecting a subset of variables as direct predictors for an outcome is a much studied problem in regression modelling and has received renewed attention in the context of high-dimensional data where some variable selection is unavoidable. It appears self-evident that best subset selection (BSS)[1–3] should be the gold standard for variable selection: Clearly, if we assume there are s direct predictors and consider *all* combinations of variables up to a subset size $k \geq s$ the true model has to be one of the candidate models, making BSS the obvious choice for variable selection. The main reason for dismissing BSS is that a naive implementation quickly becomes

computationally infeasible with larger numbers of variables[1] [4]. However, recent developments have cast a shadow on the performance of BSS even when computationally feasible. In their path-breaking work, *Bertsimas et al.* [5] have formulated BSS as a mixed integer optimization problem (MIO) pushing the boundaries for the feasible number of variables p to be in the thousands while still searching over moderate subset sizes of k. This now allows a better comparison of BSS with variable selection methods such as the popular Least absolute shrinkage estimator (Lasso) [6] or variants thereof, e.g. the adaptive Lasso [7], the Sparse-Groupe Lasso [8] or the Elastic net (Enet) and its adaptive version [9, 10]. The convex optimization nature of all these methods enables quick computation even for millions of variables making them the main methods of choice in high-dimensional settings. While these methods' theoretical and empirical performances have been studied in much detail [11–15], the question arises how they actually compare empirically to BSS in realistically high-dimensional settings.

A first extensive comparison of BSS with the Lasso and a simplified version of the relaxed Lasso[16] has been carried out by *Hastie et al.* [4]. Perhaps surprisingly, the authors find that neither BSS nor the Lasso uniformly dominate each other, and moreover that forward step-wise selection is mostly as good as BSS, while the relaxed Lasso exhibits best performance overall. An explanation may be found in the different bias-variance tradeoffs of the different appraoches. However, *Hastie et al.* [4] focus on the predictive performance of the methods, considering several popular metrics of predictive accuracy. Their results may therefore not hold up for variable selection performance since different sets of selected variables can give very similar predictions but only one set of variables is the true set of direct predictors. Another recent simulation study using the MIO formulation was carried out by *Takano et al.* [17]. The authors compared BSS with Lasso based on different optimization criteria determining the subset size k. However, in their study they only considered a low dimensional setting with $p = 100$.

In the present paper we complement the above studies by specifically evaluating the selection performance of BSS compared to established variable selection approaches, gaining further insights into the properties of all the methods. Selecting the true direct predictors of an outcome is a distinct problem from prediction and often of much substantial interest in its own right, for instance in genetics. We take advantage of the MIO formulation, which makes BSS feasible in practically relevant high-dimensional setting. While we choose a similar setting for our simulation study as *Hastie et al.* [4], we also extend their approach in several important ways: For wider applicability, we consider more complex situations than just the Toeplitz correlation structure; we also choose different positions of the direct predictors within those correlation structures. All-in-all, our simulation has 270 different parameter combinations and the results can be compared in an interactive web-app that we created [18].

All methods evaluated, here, require choosing a tuning parameter or subset size potentially affecting which and how many variables are selected. To enable a fair comparison, we choose for each method its optimal tuning parameter (or subset size)

[1]For example, with $p = 100$ and a subset size $k = 15$ there are over $2.533 \cdot 10^{17}$ possible sets.

in terms of its best achievable F1-score. This allows us to assess the best *possible* variable selection performance, separating this from the issue of choosing a tuning parameter.

The paper is organized as follows: The methods section describes the selection procedures under investigation and their theoretical properties. Subsequently, we describe the set-up of the simulation study as well as our findings. Finally, we will draw some conclusions and give an outlook onto future work in the last section.

Methods

Given a vector of responses $\mathbf{y} \in \mathbb{R}^n$, a design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, a vector of coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ and a noise vector $\epsilon \in \mathbb{R}^n$ with $\epsilon_i \sim \mathcal{N}(0, \sigma)$ for independent $i = 1, \ldots, n$, we assume the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{x}_j$, $j = 1, \ldots, p$, has been standardized such that $\sum_{i=1}^n x_{i,j} = 0$ and $n^{-1} \sum_{i=1}^n x_{i,j}^2 = 1$. We further assume $\boldsymbol{\beta}$ to be sparse in the sense that for $s = \sum_{j=1}^p I(\beta_j \neq 0)$ we have $s = \mathcal{O}(n^c)$ for $0 < c < 1$ [19, 20]. We define $\mathbf{b} = (b_1, \ldots, b_j, \ldots, b_p)^\top$ as the model vector that indicates the true model by its entries

$$b_j = \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{else.} \end{cases}$$

Since $\boldsymbol{\beta}$ has to be sparse most entries of $\mathbf{b}$ are 0. For selection we are rather interested in the entries of $\mathbf{b}$ than in the exact values of $\boldsymbol{\beta}$. Therefore, let $\hat{\boldsymbol{\beta}}$ denote an estimator of $\boldsymbol{\beta}$ that we use to construct $\hat{\mathbf{b}}$ by setting its entries as

$$\hat{b}_j = \begin{cases} 1 & \text{if } \hat{\beta}_j \neq 0 \\ 0 & \text{else.} \end{cases}$$

An estimator selects the correct variables if $\text{supp}(\hat{\mathbf{b}}) = \text{supp}(\mathbf{b})$, where $\text{supp}(\cdot)$ denotes the support of a vector. We say that an estimator or procedure is selection consistent if $\text{supp}(\hat{\mathbf{b}})$ converges in probability to the true support.

While the ordinary least squares (OLS) estimator is the best linear unbiased estimator for $\boldsymbol{\beta}$ when $n > p$ it is not useful for variable selection where the aim is to discriminate zeros from non-zeros in $\boldsymbol{\beta}$. For $\beta_j = 0$ it can be shown that the OLS estimate is $\hat{\beta}_j = \mathcal{O}\left(\sqrt{n^{-1} \log n}\right)$ [21], i.e. it does not select a model for finite n as estimated coefficients will not be exactly zero. In a high dimensional setting $p > n$ the OLS estimate is not unique [22].

For variable selection, different penalized least squares approaches have been formulated as an optimization problem of the form

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q, \tag{2}$$

where the penalty term $||\boldsymbol{\beta}||_q := \left(\sum_{j=1}^{p} |\beta_j|^q\right)^{1/q}$ denotes the L_q-norm with special case $||\boldsymbol{\beta}||_0 := \sum_{j=1}^{p} I\left(\beta_j \neq 0\right)$. The tuning parameter $\lambda \geq 0$ controls the strength of the penalty where for $q < 2$ a higher λ shrinks $\hat{\beta}_j$ stronger towards 0 and fewer variables are selected. Choosing λ can be based on criteria like AIC, BIC, cross-validation or stability procedures [23–26] each with different goals and different assumptions [27–29]. In the following we will focus on some of the most prominent penalization approaches.

Best subset selection

Using the L_0-norm in (2) is known as Best Subset Selection (BSS) and can be formulated as the following discrete optimization problem

$$\hat{\boldsymbol{\beta}}_{BSS} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \qquad \text{s.t. } ||\boldsymbol{\beta}||_0 \leq k \tag{3}$$

with $k \in \mathbb{N}$ determining the maximal number of non-zeros in $\hat{\boldsymbol{\beta}}$. *Zhang et al.* [30] showed that if a uniform signal strength condition for the smallest true predictor in $\boldsymbol{\beta}$ holds the BSS can achieve selection consistency with respect to $\mathbf{b}$. *Shen et al.* [31] defined a degree of separation that describes how difficult it is to discriminate the true model from all other models in terms of the projection of $\mathbf{y}$ based on $\hat{\boldsymbol{\beta}}$. As a necessary condition for a L_0-norm based penalty approach to be selection consistent they showed that the degree of separation has to be larger than a threshold that is a function of p, n and σ^2.

The minimization of (3) is known to be NP-hard [32, 33] and state-of-the-art algorithms have been capable of solving BSS problems in a feasible amount of time only if $p < 50$. Recently Bertsimas et al. [5] reformulated (3) as a mixed integer optimization (MIO) problem

$$\begin{aligned} &\arg\min_{\boldsymbol{\beta},\mathbf{z}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 \\ &\quad \text{s.t. } (\beta_i, 1 - z_i) \ : \ \text{SOS-1}, \qquad i = 1, \ldots, p \\ &\qquad z_i \in \{0, 1\}, \qquad i = 1, \ldots, p \\ &\qquad \sum_{i=1}^{p} z_i \leq k. \end{aligned} \tag{4}$$

where SOS-1 denotes a Specially Ordered Set of Type 1, i.e. at most one element of $(\beta_i, 1 - z_i)$ can be non-zero. The authors showed that this reformulation guarantees optimality in the sense of (3). Due to efficient MIO solvers like Gurobi [34] problem (4) can be solved in minutes even when p is in the 1000s, n in the 100s and a moderate value k is selected. However, certifying the optimality of the solution can take much more time. For example the Gurobi solver uses a lower an upper bound criterium to find a solution where the convergence rate of the lower bound criterium is much faster [4].

Forward step-wise selection

While BSS is limited through its NP-hard nature, stepwise selection is a popular alternative. It gradually adds (removes) variables to (from) a model based on some

criterion of model fit. Due to this greedy strategy these algorithms are computationally less challenging with complexity $\mathcal{O}(p^2)$. However, stepwise selection approaches are known to have numerous drawbacks: they result in unstable final models that are sensitive to small changes in the data [35–37] and they are only locally optimal and often miss direct predictors while selecting irrelevant variables [38, 39]. Moreover, inference is problematic as they usually do not account for multiple testing issues [37, 40]. Despite these drawbacks we will consider here forward stepwise selection (FSS) [41, 42] in our simulation study because it can be interpreted as greedy heuristic of BSS [4]. It is defined as an iterative algorithm and starts with an empty active set model $A_0 = \emptyset$ and $\hat{\boldsymbol{\beta}}^{(k)}_{A_0} = \mathbf{0}$. At each step $t = 1, \ldots, k$ the variable j_t is selected that maximizes

$$\arg \max_{j_t \notin A_{t-1}} \frac{\mathbf{x}_{j_t}^\top (\mathbf{I} - \mathbf{P}_{A_{t-1}})\mathbf{y}}{||(\mathbf{I} - \mathbf{P}_{A_{t-1}})\mathbf{x}_{j_t}||_2}$$

where $P_{A_{t-1}}$ denotes the projection of $\mathbf{y}$ onto the column space of $\mathbf{X}_{A_{t-1}}$. Given j_t, the active set is updated as $A_t = A_{t-1} \cup \{j_t\}$ and used to estimate

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(t)}_{A_t} &= \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}_{\mathbf{A_t}}\boldsymbol{\beta}||_2^2, \\ \hat{\boldsymbol{\beta}}^{(t)}_{\setminus\{A_t\}} &= 0 \end{aligned}$$

where $\setminus\{A_t\}$ denotes the set of not selected predictors at step t.

Lasso and Elastic net

The least absolute shrinkage and selection operator (Lasso) [6] uses an L_1-norm as penalty term in (2) and shrinks all estimated coefficients by an absolute value towards zero. For a sufficiently large tuning parameter λ the estimated coefficients are exactly zero, hence, the Lasso performs variable selection. The total number of zero coefficients is controlled by the λ where larger values will result in sparser models. Although the Lasso can be combined with fast algorithms so that problems with p in the $10,000$'s can easily be solved, i.e. for settings for which BSS with MIO can no longer be applied, it also has several drawbacks. Firstly, it only allows up to n non-zero regression coefficients which can be a limiting factor if $n \ll p$ [42]. Secondly, if irrelevant variables are highly correlated with direct predictors of the outcome $\mathbf{y}$ the Lasso selects almost arbitrarily of those true and false variables and is known not to consistent, not even for the sign of the coefficient. Further, if there is a high pairwise correlation within a set of variables and they all are true direct predictors for $\mathbf{y}$ the Lasso tends to select only one of these variables [9, 43]. Thus, it can not guarantee consistent variable selection [20]. To address some of these drawbacks, different modifications of the Lasso have been proposed. They rely on a priori knowledge about the data generating process or the functional relationship between variables [7, 8, 16, 44–46]. An alternative is the Elastic net (Enet) [9] which can be formulated as a weighted combination of the Lasso and the an additional L_2-penalization (Ridge)

$$\hat{\boldsymbol{\beta}}_{Enet} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \alpha\lambda||\boldsymbol{\beta}||_1 + (1-\alpha)\lambda||\boldsymbol{\beta}||_2. \quad (5)$$

The second tuning parameter $0 \leq \alpha \leq 1$ controls the weighting between the L_1- and L_2-penalty. Here the L_1-penalty induces a Lasso-type variable selection while the L_2-penalty helps with highly correlated variables by increasing the diagonal entries of the covariance matrix $\mathbf{X}'\mathbf{X}$. The latter guarantees a positive-definite covariance matrix so that it is possible for all p estimated coefficients to be non-zero. More importantly, the L_2-penalty can be interpreted as an artificial decorrelation of the variables making it easier to jointly select highly correlated variables if they are all direct predictors for $\mathbf{y}$ [9]. However, the double shrinkage of the Enet increases the bias of the estimators more than Lasso or Ridge alone and a rescaling of the estimators, that does not affect the number of estimated non-zeros, has been suggested [9] .

Simulation study

Simulation design and evaluation

To evaluate the performance of BSS, FSS, Lasso and Enet for variable selection we simulated data from a linear model (1) with $\mathbf{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Since real world applications of variable selection often have a small signal-to-noise ratio τ we followed [4] and set $\sigma^2 = \frac{\boldsymbol{\beta}^\top \mathbf{\Sigma} \boldsymbol{\beta}}{\tau}$ with $0.05 \leq \tau \leq 6$.

For a low-dimensional setting we chose $n = 1000$ and $p = 100$ and for a high dimensional setting $n = 100$ and $p = 1000$. Additionally, we considered an intermediate setting with $n = 500$ and $n = 500$. To assess the role of correlation between predictors we used an uncorrelated structure and the standard Toeplitz structure. The latter was created by setting the pairwise correlation between two variables $x_{\cdot,u}$ and $x_{\cdot,v}$ for $u, v = 1, \ldots, p$ as $\rho^{|u-v|}$ with $\rho \in \{0.35, 0.7\}$. However, although the Toeplitz structure is a popular correlation structure in simulations, it can be implausible in some applications. For example, in genetic epidemiology it is often reasonable to assume that genes within a functional group are correlated with each other, but that they are nearly independent of genes from other functional groups. Hence, we simulated data with correlations following a block structure for which variables were grouped into blocks of size 10 with pairwise correlations of $\rho \in \{0.35, 0.7\}$ within blocks and $\rho = 0$ between variables of different blocks. The number of direct predictors was set to $s = 10$ in all settings and their position was either consecutive or equally spaced along the the sequence of variable. For the consecutive positioning we set the first ten coefficients to be non-zeros while for the equally spaced positioning we set every tenth coefficient to be non-zero (see Figure 1). In all scenarios a non-zero direct predictor was set to $\beta_j = 1$. Overall we investigated 270 different scenarios and each one was repeated 100 times. As performance measure we used the F1-score

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

where $P = \frac{TP}{TP+FP}$ is precision and $R = \frac{TP}{TP+FN}$is recall (and TP is the number of true positives, FP of false positives and FN of false negatives).

Selection of tuning parameter

All methods considered here rely either on choosing a tuning parameter λ or a maximum set size k *a priori*. The Enet even requires setting two tuning parameters.

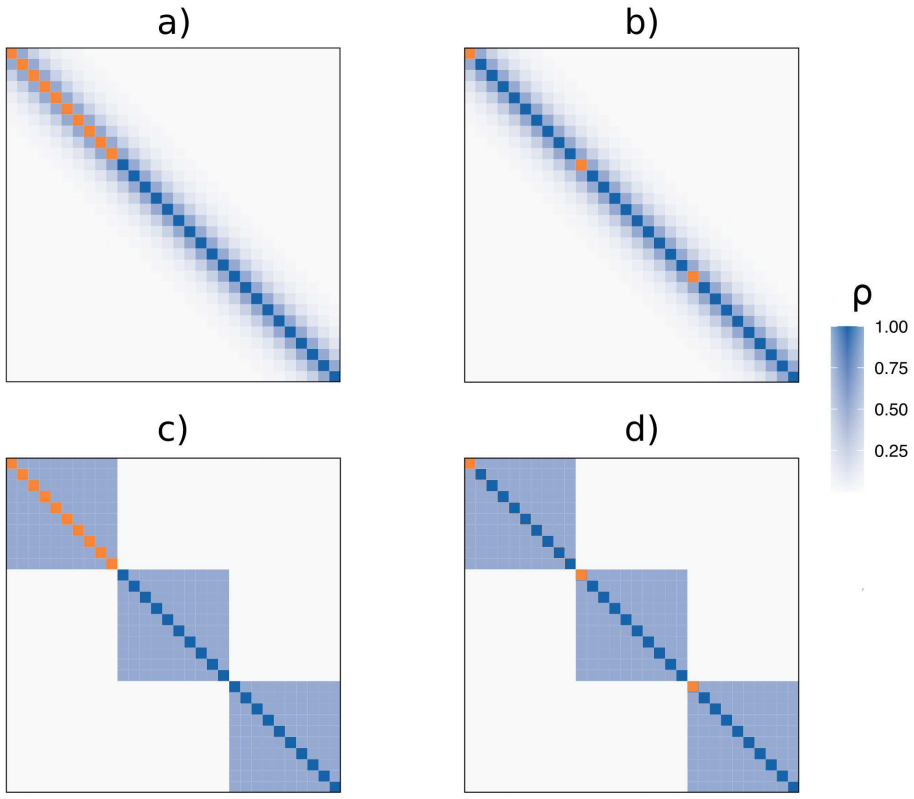

Figure 1 Schematic representation of the different correlation structures and the positioning of the non-zeros coefficients (orange). The Toeplitz structure with all ten non-zeros in consecutive positioning (a) and equally spaced positioning (b) has a gradually decreasing correlation between all variables. The second row shows a block structure where all coefficients of one block are set to non-zero (c) or only one coefficient of each block is set as non-zero for 10 blocks (d).The correlation between variables within one block is always the same while the correlation between variables of different blocks is set to zero.

Obviously, the performance of each variable selection methods depends the choice of these tuning parameter(s) or subset size. Since we are interested in the best *possible* performance of each method regarding variable selection we used a grid of tuning parameters λ and α for Lasso/Enet, i.e. $\alpha = 0.1, 0.2, \ldots, 0.9$ and 1000 values for λ where the largest λ returns an empty model and the smallest λ a full model. For BSS and FSS we used subset sizes of $k = 1, \ldots, 20$ which means that for $k = 10$ both methods had the chance to find the true model. In a final step, for each method to be compared, we only chose those tuning values / set sizes that gave the highest F1 score in the considered setting.

Although MIO much accelerates BSS for our high dimensional settings, it can still run for hours making an extensive comparison infeasible without a time limit. Following the suggestion of other authors [4, 5], we set the time limit to 3 minutes, which is sufficient for the Gurobi solver to find a solution. However, certifying for optimality can take much longer. Because this could disadvantage BSS compared to the other methods we investigate the number of certified optimality BSS solutions and its impact on the F1 score.

Results

Surprisingly, BSS reliably outperformed the other methods only in settings with high signal-to-noise ratio and when the variables were uncorrelated. Even in a *low dimensional* setting, with the number of observations ten times the number of variables, the selection performance of BSS drops dramatically if the true predictors are moderately correlated. In those cases BSS is even outperformed by the Lasso which is known to be inconsistent for variable selection when the true predictors are highly correlated. Interestingly, the much simpler FSS achieves a similar performance as BSS in almost all settings, in some even slightly better.

To give a more detailed impression of the BSS performance we will report the results of four specific settings regarding dimensionality, non-zero position as well as correlation structure and hight. The remaining results are shown in the supplement "Additional file 1" or interactive web app [18] and support our main conclusions. Note, for the Enet we only show results with $\alpha \in \{0.1, 0.5, 0.9\}$ representing a mostly Ridge-weighted, a balanced as well as a mostly Lasso-weighted Enet.

Variable selection performance

In general all methods perform better under a Toeplitz structure compared to a block structure. This seems plausible since the correlations under the Toeplitz correlation structure are weaker than within blocks. For the high dimensional block setting with equally spaced non-zeros, $\rho = 0.35$ and low τ, all methods have a relatively small best F1-score (see Figure 2). FSS and BSS show nearly identical results and only outperform Lasso and Enet with large signal-to-noise ratios with respect to the F1-score ($\tau \geq 3.52$) and precision ($\tau \geq 0.71$). Lasso and Enet show better recall values on average except for $\tau \geq 3.52$ while the variability is high for $\tau \leq 0.14$.

Figure 3 shows the results for a high dimensional Toeplitz structure setting with consecutive non-zeros. The methods exhibit large differences: for $\tau \geq 2.07$ the Ridge-weighted Enet ($\alpha = 0.1$) achieves mostly a very high F1-score of 1, benefiting clearly from the decorrelation. In comparison, the other methods cannot cope with highly correlated direct predictors as seen from the low recall values and ensuing low F1-score values of the Lasso and the weak performances of BSS and FSS. Figure 3 shows that the recall of BSS and FSS slightly benefits from an increase in τ.

In the low-dimensional block setting with $\rho = 0.7$ and equally spaced non-zeros BSS and FSS barely outperform Lasso and Enet for most τ. Figure 4 shows that this is mainly due to the relatively high precision. However, when the non-zeros are consecutive the performance of FSS and BSS decreases drastically. Even in the low dimensional case the signal-to-noise ratio has to be very large ($\tau = 6$) to achieve comparable results to Lasso and Enet (see Figure 5). In this cases the F1-scores of FSS and BSS are dominated by their poor recall.

Certified runs

Although we needed to set a time limit for the Gurobi solver (which we chose so that computation time would stay below 56 days), the poor performance of BSS cannot be explained by non-certified solutions. Figure 6 shows the number of certified and non-certified optimal solutions for the low-dimensional setting with consecutive non-zeros of figure 5. For $\tau \geq 3.52$ the Gurobi solver certified all solutions and at least 80

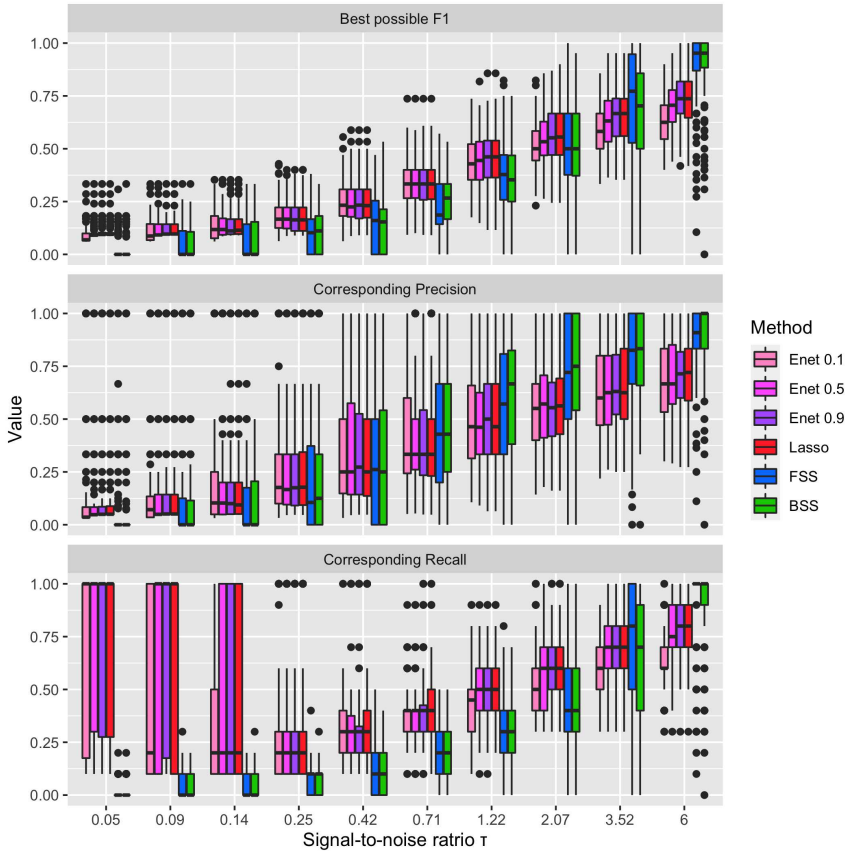

Figure 2 High-dimensional setting ($p = 1000$, $n = 100$, $\rho = 0.35$). Boxplots of best possible F1-scores and corresponding precision and recall values for a block correlation setting with equally spaced non-zero coefficients.

out of 100 runs have been certified as optimal for smaller τ. This suggests that BSS could not achieve a much better recovery of supp($\mathbf{b}$) in these settings. Conversely, an uncertified solution does not imply that BSS performs poorly. For example figure 7 shows the number of certified results for an intermediate Toeplitz setting ($p = 500$, $n = 500$, $\rho = 0.35$) with equally spaced true predictors. No run in this scenario could be certified for $\tau = 0.71$ but BSS still achieved an F1 score of 1 in 98 out of 100 runs (see [18] and supplement "Additional file 1"). This means the Gurobi solver found an F1-optimal solution that is also the best possible solution but could not certify its optimality within 3 minutes.

Conclusion and Discussion

We carried out an extensive simulation study to compare the performance of BSS and its competitors regarding variable selection. We investigated a broad range of parameter constellations as well as different and maybe more realistic correlation structures than previous works have done. Our results show that the Enet and the Lasso outperform BSS and FSS in most scenarios. This was unexpected since due to the maximal subset size $k = 20$ BSS must have considered the true model with $s = 10$ as a candidate model in every run. Hence, we would have expect BSS

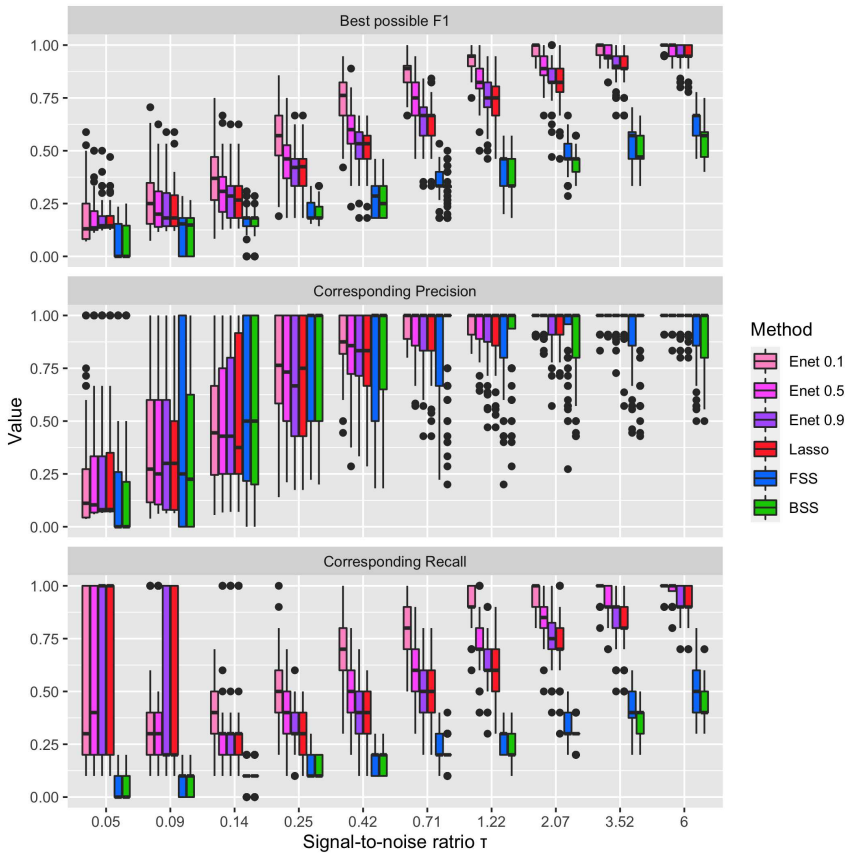

Figure 3 High-dimensional setting ($p = 1000$, $n = 100$, $\rho = 0.7$). Boxplots of best possible F1-scores and corresponding precision and recall values for a Toeplitz correlation setting with consecutive non-zero coefficients.

to perform similarly or better than the other methods. Perhaps more surprisingly, we found that BSS cannot achieve a good variable selection performance in low dimensional settings if there is a moderate to large correlation ($\rho \geq 0.35$) of the true predictors. In this case, BSS can only be 'saved' by a relative high signal-to-noise ratio which will be implausible in many practical situations. We also argue that the poor performance of BSS can not be explained by non-certified runs alone.

Based on our empirical results, using a Ridge-weighted Enet seems a good choice for settings with correlated predictors. We can recommend an L_0-norm penalization only in situations where a high signal-to-noise ratio and (nearly) uncorrelated true predictors are plausible. However, in view of the NP-hard nature of BSS and its performance being matched almost exactly by FSS, we see no benefit in choosing BSS over FSS.

Our complete results can be accessed via a web app [18] and via the supplement "Additional file 1". However, it has to be kept in mind that our simulation study was designed to evaluate variable selection methods, and not to assess the criteria for selecting the tuning parameters or subset size.

In future research, it might be promising to combine BSS or FSS with approaches like Lasso or Enet so as to preselect or decorrelate covariates. Hence, further insights

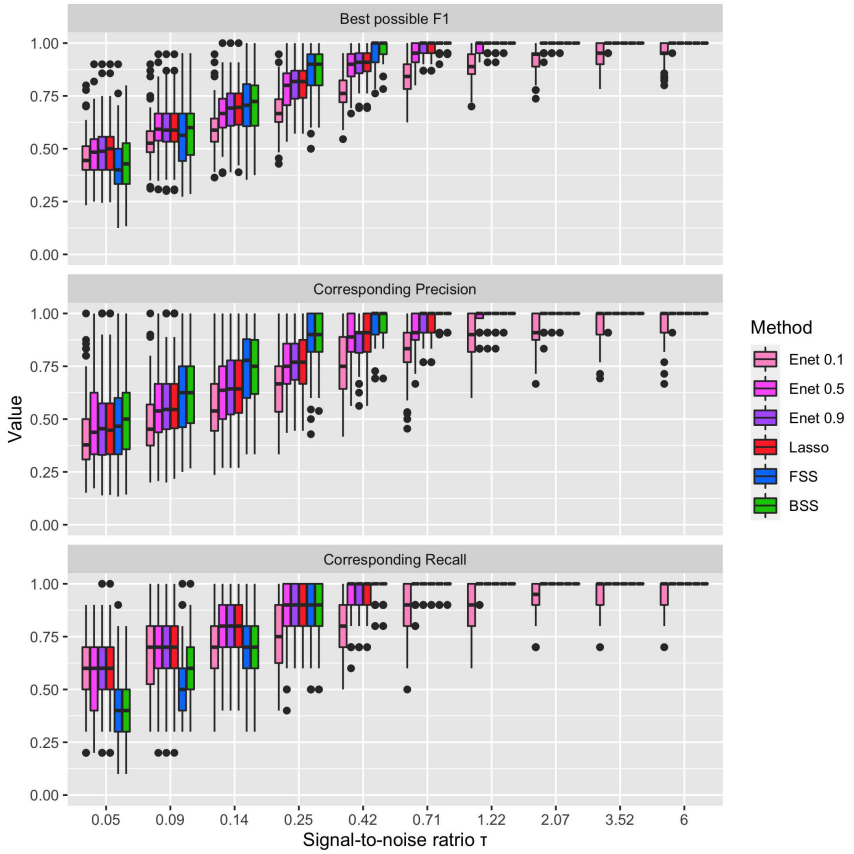

Figure 4 Low-dimensional setting ($p = 100$, $n = 1000$, $\rho = 0.7$). Boxplots of best possible F1-scores and corresponding precision and recall values for a block-structured correlation setting with equally spaced non-zero coefficients.

on the role of the correlation structure for variable selection performance of BSS are desirable.

Abbreviations
BSS: Best subset selection
Enet: Elastic net
FSS: Forward stepwise selection
Lasso: Least absolute shrinkage operator
MIO: Mixed integer optimization

Competing interests
The authors declare that they have no competing interests.

Author's contributions
L.D., M.H., V.D. and R.F. conceived of the presented idea. The simulation set-up was defined by M.H., L.D. and V.D. The analysis and interpretation of the results were done by L.D., M.H., V.D. and R.F. The implementation was done by L.D. and M.H. The writing was done by M.H. and V.D.

Acknowledgements
We thank the Deutsche Forschungsgemeinschaft (DFG) for funding this research.

Funding
This work was funded by the Deutsche Forschungsgemeinschaft (DFG), project number FO 1045/2-1.

Availability of data and materials
All data of this simulation study can be generated by the R-code under github.com/bips-hb/bscomparison and github.com/bips-hb/simsham. All results of the simualtion can be accessed under https://bestsubset.bips.eu.

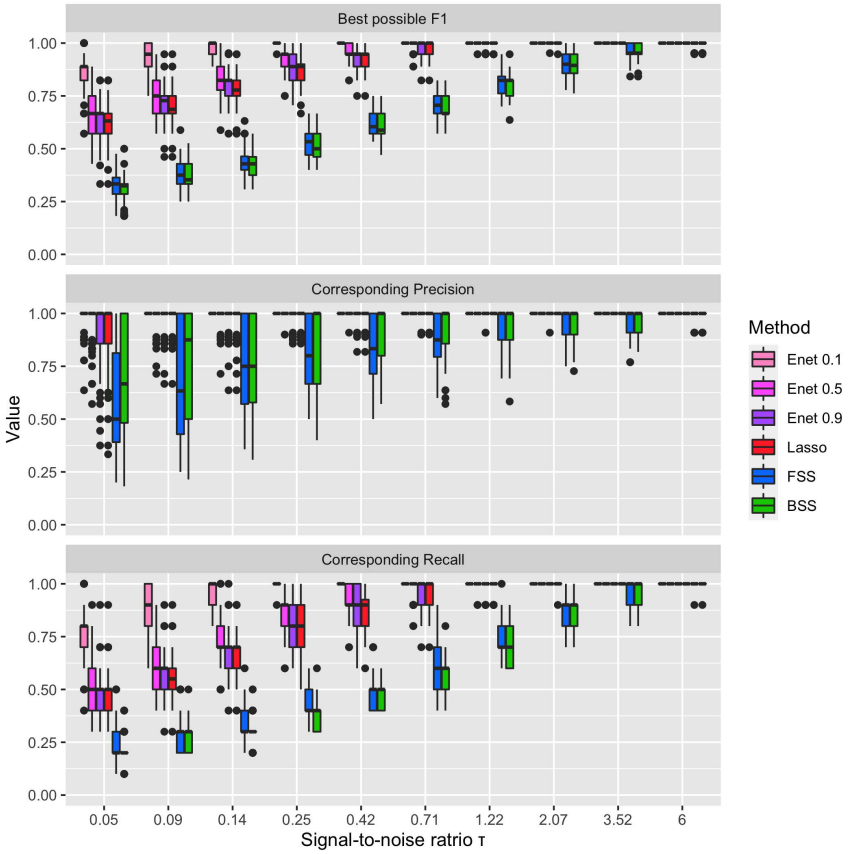

Figure 5 Low-dimensional setting ($p = 100$, $n = 1000$, $\rho = 0.7$). Boxplots of best possible F1-scores and corresponding precision and recall values for a block-structured correlation setting with consecutive non-zero coefficients.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

References

1. Beale, E.M.L., Kendall, M.G., Mann, D.W.: The discarding of variables in multivariate analysis. Biometrika **54**(3/4), 357–366 (1967)
2. Hocking, R.R., Leslie, R.N.: Selection of the best subset in regression analysis. Technometrics **9**(4), 531–540 (1967)
3. Garside, M.J.: Some Computational Procedures for the Best Subset Problem. Journal of the Royal Statistical Society Series C **20**(1), 8–15 (1971)
4. Hastie, T., Tibshirani, R., Tibshirani, R.J.: Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso (2017). 1707.08692
5. Bertsimas, D., King, A., Mazumder, R.: Best subset selection via a modern optimization lens. The Annals of Statistics **44**(2), 813–852 (2016)
6. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society (Series B) **58**, 267–288 (1996)
7. Zou, H.: The adaptive lasso and its oracle properties. Journal of the American Statistical Association **101**(476), 1418–1429 (2006)
8. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: A sparse-group lasso. Journal of Computational and Graphical Statistics **22**(2), 231–245 (2013)
9. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B **67**, 301–320 (2005)
10. Zou, H., Zhang, H.H.: On the adaptive elastic-net with a diverging number of parameters. Ann. Statist. **37**(4), 1733–1751 (2009)

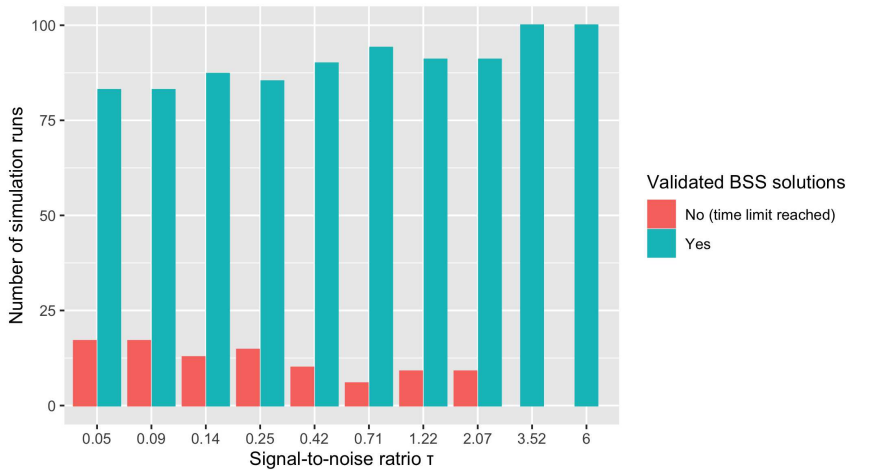

Figure 6 Number of simulation runs with certified optimal solution found by the Gurobi solver for a block-structured low-dimensional setting with $\rho = 0.7$ and consecutive true predictors.

11. Chun, H., Keleş, S.: Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Series B (Statistical Methodology) **72**(1), 3–25 (2010)
12. Houwelingen, H.C., Sauerbrei, W.: Cross-validation, shrinkage and variable selection in linear regression revisited. Open Journal of Statistics **3**, 79–102 (2013)
13. Sanchez-Pinto, L., Venable, L., Fahrenbach, J., Churpek, M.: Comparison of variable selection methods for clinical predictive modeling. International Journal of Medical Informatics **116** (2018)
14. Yu, X., Ge, H., Lu, D., Zhang, M., Lai, Z., Yao, R.: Comparative study on variable selection approaches in establishment of remote sensing model for forest biomass estimation. Remote Sensing **11**(12) (2019)
15. Lima, E., Davies, P., Kaler, J., Lovatt, F., Green, M.: Variable selection for inferential models with relatively high-dimensional data: Between method heterogeneity and covariate stability as adjuncts to robust selection. Scientific Reports **10** (2020)
16. Meinshausen, N.: Relaxed lasso. Computational Statistics and Data Analysis **52**, 374–393 (2007)
17. Takano, Y., Miyashiro, R.: Best subset selection via cross-validation criterion. TOP **28**(2), 475–488 (2020)
18. Hanke, M., Dijkstra, L., Foraita, R., Didelez, V.: Simulation Results for BSS, FSS, Lasso and Enet. https://bestsubset.bips.eu
19. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the lasso. The Annals of Statistics **34**(3), 1436–1462 (2006)
20. Zhao, P., Yu, B.: On model selection consistency of lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)
21. Horowitz, J.L.: Variable selection and estimation in high-dimensional models. Canadian Journal of Economics/Revue canadienne d'économique **48**(2), 389–407 (2015)
22. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman & Hall/CRC, Boca Raton, Florida (2015)
23. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (eds.) Selected Papers of Hirotugu Akaike vol. 1, 1st edn., pp. 199–213. Springer, New York, NY (1998)
24. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**, 461–464 (1978)
25. Picard, R.R., Cook, R.D.: Cross-validation of regression models. Journal of the American Statistical Association **79**(387), 575–583 (1984)
26. Liu, H., Roeder, K., Wasserman, L.A.: Stability approach to regularization selection (stars) for high dimensional graphical models. Advances in neural information processing systems **24 2**, 1432–1440 (2010)
27. Shao, J.: An asymptotic theory for linear model selection. Statistica Sinica **7**(2), 221–242 (1997)
28. Yang, Y.: Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. Biometrika **92**(4), 937–950 (2005)
29. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Statistics Surveys **4**, 40–79 (2010)
30. Zhang, C.-H., Zhang, T.: A general theory of concave regularization for high-dimensional sparse estimation problems. Statist. Sci. **27**(4), 576–593 (2012). doi:10.1214/12-STS399
31. Shen, X., Pan, W., Zhu, Y., Zhou, H.: On constrained and regularized high-dimensional regression. Annals of the Institute of Statistical Mathematics **65(5)**, 807–832 (2013)
32. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM J. Comput. **24**(2), 227–234 (1995)
33. Barron, A., Birgé, L., Massart, P.: Risk bounds for model selection via penalization. Probability Theory and Related Fields **113**(3), 301–413 (1999)
34. Gurobi Optimization, L.: Gurobi Optimizer Reference Manual (2021). http://www.gurobi.com
35. James, F.C., McCulloch, C.E.: Multivariate analysis in ecology and systematics: Panacea or pandora's box?

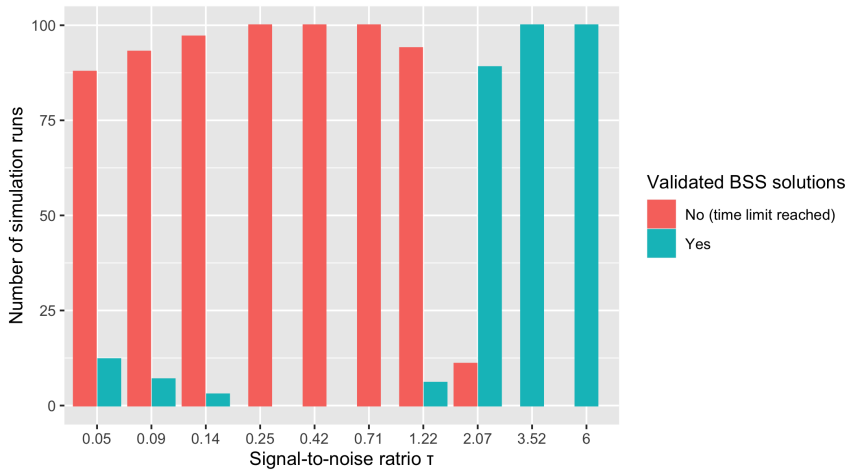

Figure 7 Number of simulation runs with certified optimal solution found by the Gurobi solver for a Toeplitz-structured intermediate-dimensional setting with $\rho = 0.35$ and equally spaced true predictors.

Annual Review of Ecology, Evolution, and Systematics **21**, 129–166 (1990)
36. Breiman, L.: Bagging predictors. Machine Learning **24**, 123–140 (1996)
37. Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P.: Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology **75**(5), 1182–1189 (2006)
38. Derksen, S., Keselman, H.J.: Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology **45**(2), 265–282 (1992)
39. Smith, G.: Step away from stepwise. Journal of Big Data **5**(1), 32 (2018)
40. Mundry, R., Nunn, C.: Stepwise model fitting and statistical inference: Turning noise into signal pollution. The American Naturalist **173**(1), 119–123 (2009)
41. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist. **32**(2), 407–499 (2004)
42. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd edn. Springer, New York (2009)
43. Xu, H., Caramanis, C., Mannor, S.: Sparse algorithms are not stable: A no-free-lunch theorem. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(1), 187–193 (2012)
44. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society Series B, 91–108 (2005)
45. Friedman, J., Hastie, T., Tibshirani, R.: A note on the group lasso and a sparse group lasso (2010). 1001.0736
46. Alaíz, C.M., Barbero, Á., Dorronsoro, J.R.: Group fused lasso. In: Mladenov, V., Koprinkova-Hristova, P., Palm, G., Villa, A.E.P., Appollini, B., Kasabov, N. (eds.) Artificial Neural Networks and Machine Learning – ICANN 2013, pp. 66–73. Springer, Berlin, Heidelberg (2013)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Choosingthebestsubset.bib