Computational Statistics and Data Analysis 71 (2014) 725–742

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Using random subspace method for prediction and variable importance assessment in linear regression

Jan Mielniczuk [a,b,*], Paweł Teisseyre [a]

[a] Institute of Computer Science, Polish Academy of Sciences, Poland

[b] Warsaw University of Technology, Faculty of Mathematics and Information Science, Poland

ARTICLE INFO

Article history:
Received 16 December 2011
Received in revised form 24 September 2012
Accepted 28 September 2012
Available online 9 October 2012

Keywords:
Random subspace method
High-dimensional model selection
Prediction
Variable importance
Positive selection rate
False discovery rate

ABSTRACT

A random subset method (RSM) with a new weighting scheme is proposed and investigated for linear regression with a large number of features. Weights of variables are defined as averages of squared values of pertaining t-statistics over fitted models with randomly chosen features. It is argued that such weighting is advisable as it incorporates two factors: a measure of importance of the variable within the considered model and a measure of goodness-of-fit of the model itself. Asymptotic weights assigned by such a scheme are determined as well as assumptions under which the method leads to consistent choice of significant variables in the model. Numerical experiments indicate that the proposed method behaves promisingly when its prediction errors are compared with errors of penalty-based methods such as the lasso and it has much smaller false discovery rate than the other methods considered.

1. Introduction

Prediction problem with a high dimensional feature space is one of the most challenging tasks of contemporary applied statistics. There is a growing number of domains nowadays that produce data with a large number of features, while the number of observations is limited. Examples include microarray datasets that measure genes activity, Quantitative Trait Loci (QTL) data, drug design datasets, high-resolution images and high-frequency financial data among others. For examples and discussion, see e.g., Donoho (2000). An important and intensively studied line of research is focused on regularization, or penalty-based methods (cf. e.g., Tibshirani, 1996; Zou and Hastie, 2005). Another important approach is a method of dimensionality reduction based on the so called sure independence screening proposed by Fan and Lv (2008). Recently, Bühlmann et al. (2010) have introduced a novel, computationally feasible method relying on a certain hierarchical testing algorithm. There are also approaches using information criteria modified to the high-dimensional setup; see e.g., Frommlet et al. (2012). In this paper, we propose a different approach based on the random subset method (RSM).

In the RSM a random subset m of features having cardinality $|m|$ smaller than a number of potentially useful regressors M is chosen and the problem is solved with the reduced feature space of the selected predictors. Features under consideration are assigned weights based on their performance in the constructed solution. The selection of a random subset of features and model fitting is executed B times and a cumulative weight of a feature is calculated based on its relevance in all models where it is used. The cumulative weights (or scores) thus correspond to relative importance of variables in the considered problem. The variables are then ordered according to the assigned weights. The ordering is essential in a construction of a final model, which can be e.g. based on a predetermined number of the most significant predictors or obtained by a selection

* Correspondence to: Jana Kazimierza 5, 01–248, Warsaw, Poland. Tel.: +48 22 3800500; fax: +48 22 3800510.
E-mail addresses: miel@ipipan.waw.pl (J. Mielniczuk), teisseyrep@ipipan.waw.pl (P. Teisseyre).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.09.018

method applied to the hierarchical list of models given by the ordering. By choosing $|m|$ much smaller than M the problem of overfitting is circumvented. Note that in an extreme case when $|m| = 1$ the scores correspond to the individual performance of variables.

The procedure was proposed by Ho (1998) for classifying objects and independently by Breiman (2001) for the case when a considered prediction method was either a classification or a regression tree. Breiman's approach leads to a construction of a random forest. There, a score of a feature corresponds to the difference of prediction errors averaged over trees which used this feature and its analogue for which values of the variable are randomly permuted. For the important developments, see also Lai et al. (2006) and Draminski et al. (2008).

The RSM method belongs to the category of wrappers in the sense that feature selection is 'wrapped around' building a prediction method i.e. it is inherent part of its construction. Here, all variables are ranked first based on their averaged performance in small fitted models and then selection of variables is performed for ensuing hierarchical family of models with the use of cross-validation or independent test sample. Different group, called filters, includes methods for which the feature selection method is not related to the construction of a prediction tool. Such methods can perform ranking and variable selection simultaneously; for a representative example see e.g., Stoppiglia et al. (2003). It should be stressed however, that since ranking of variables in the RSM is based on fitting small linear models the method does not impose any conditions on the number of candidate variables M. In classification and regression it is proved to be an effective way to avoid pitfall of curse of dimensionality in situations when the number of features M is comparable or even significantly larger than the sample size n.

A related problem, which is also addressed by the RSM, is assignment of weights to the variables in such a way that their magnitudes would correspond to variables' usefulness in the prediction. Note that this problem is different from the explanation problem in which we try to determine significant variables in the 'true' model, that is a model which fits the data well. A variable may be important for prediction although it does not belong to the set of significant features even in the ideal case when the data conform to a certain model like a linear model (1). The problem of assigning scores to features which reflect their importance in prediction when the number of features is small compared to the sample size is also an important line of research; cf. Grömping (2007) for comprehensive review. Here, important development is the method proposed by Lindemann, Merenda and Gold (`lmg`); see Lindemann et al. (1980) and also Chevan and Sutherland (1991). In this approach, the score of variable x equals an average over all permutations r of $\Delta R^2_{x,r}$, where $\Delta R^2_{x,r}$ is an increase of coefficient of determination R^2 due to adding x to the list of active variables ordered by r (see in Section 4 for a formal definition). The method was further developed by Feldman (1999), who considered data dependent weights with their magnitude corresponding to the goodness-of-fit of the ordering. Note however, that for large M both approaches become extremely computationally intensive and they break down in the case of linear model fitting when M is larger than n. We also mention the weight assignment method based on Multivariate Adaptive Regression Splines (MARS) discussed in Section 4.2.

The aim of the paper is twofold. First, for a linear regression we introduce a new scheme of assigning scores to variables. In our approach variables in a randomly chosen subset are assigned weights equal the squared values of the respective t-statistics in the pertaining fitted model. We argue in Section 2 that this is an intuitively sound choice of weights as Eq. (3) indicates that the square of t-statistic is a product of two factors, one of which corresponds to the importance of variable within the model and the second to the importance of the model itself. Second, we investigate models based on the ordered features according to the proposed weighting and study their prediction strength by means of simulations. We establish some formal properties of the proposed scheme, namely we determine the form of asymptotic ordering of the variables when the subset is fixed (Theorem 1) and we establish the asymptotic form of weights assigned by the RSM (Theorems 2 and 3). In the case of fixed subset m and random regressors the ordering is asymptotically equivalent to that given by the multiple correlation coefficients of y and variables in m with a single consecutive variable dropped. This is an extension of Zheng and Loh (1995) result, who have shown that in the case when m contains all relevant variables the obtained ordering is such that the relevant variables precede all spurious ones.

The prediction accuracy of the RSM based approach is compared with that of the lasso by means of numerical experiments and its performance appears to be promising, especially when there are many potential strongly dependent regressors. It turns out that in the considered examples false discovery rate for the RSM is much smaller than for the other methods whereas positive selection rate for all of them is comparable and close to 1. Also, we compared the pertaining method of weight assignment with Breiman's measures and a weight assignment based on MARS.

The paper is structured as follows. Properties of the t-based ordering for a fixed subset of regressors are studied in Section 2 along with the examples in which the explicit conditions are given for it to be the correct one. Section 3 introduces the random subspace method and states the results for considered weighting scheme and Section 4 summarizes the outcomes of numerical experiments. Proofs of the results are relegated to the Appendix.

We define now a formal setup of the paper. Assume that observed data have the form $(\mathbf{Y}, \mathbf{X})$, where $\mathbf{Y} = \mathbf{Y}_n$ is an $n \times 1$ vector of n responses which variability we would like to explain and $\mathbf{X} = \mathbf{X}_n$ is an $n \times M$ design matrix consisting of vectors of M potential regressors collected from n objects. Responses are related to regressors by means of the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is an unobservable vector of errors, assumed to have $N(0, \sigma^2\mathbf{I})$ distribution. Vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)'$ is an unknown vector of parameters. We consider two scenarios: the case of deterministic and random $\mathbf{X}$. In the latter case rows of $\mathbf{X}_n$ constitute n independent realizations of M-dimensional random variable $\mathbf{x}$ and a vector $\mathbf{Y}$

consists of n realizations of a random variable $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. A distribution of $\mathbf{x} = (x_1, \ldots, x_M)'$ may be arbitrary, in particular the distribution of its first coordinate may be point mass at 1 corresponding to the linear model with an intercept included. Number of potential predictors $M = M_n$ may depend on n. Our results will either concern the case when M remains fixed (Theorem 2 and Corollary 3) or changes with the sample size (Theorems 1, 3 and Corollary 4). In particular M may be much larger than n, compare e.g. condition (ii) of Theorem 3, where it is only assumed that $\log M_n = o(n)$.

Suppose that some covariates are unrelated to the prediction of $\mathbf{Y}$, so that the corresponding coefficients β_i are zero. Model containing all relevant variables, i.e. those pertaining to nonzero β_i, will be called a true model. The minimal true model i.e. such that it pertains only to relevant variables will be denoted by t and $|t|$ will be the number of nonzero coefficients. It is assumed that $t \subset \{1, 2, \ldots, M\}$ and t does not change with n.

2. Variable importance assessment using t-statistics

In this section, we discuss rationale for using a squared value of t-statistic as a measure of variable importance in linear regression in a general case when a considered model may be misspecified. This is an interesting issue as it is intuitively clear that when e.g. the most important feature is mistakenly dropped from the model then a spurious feature highly correlated with it may have larger value of t-statistic than other true predictors. We study the problem in Theorem 1 which states the conditions under which such situation cannot occur. In particular, it follows from Corollary 2 that when variables are asymptotically uncorrelated the weighting will reflect the correct ordering of variables in the sense that all variables pertaining to the minimal true model will have larger weights than spurious ones.

Consider a submodel m of model (1) containing $|m|$ variables $i_1, \ldots, i_{|m|}$, where $|m|$ is a fixed integer such that $|m| < \min(M, n)$. In the following m will either denote a subset $\{i_1, \ldots, i_{|m|}\}$ or a model corresponding to this subset. Submatrix of $\mathbf{X}$ consisting of columns corresponding to model m will be denoted by $\mathbf{X}_m$. Analogously $\mathbf{x}_m$ is a subvector of $\mathbf{x}$ consisting of coordinates corresponding to m and x_i is i-th coordinate of $\mathbf{x}$. Model m with i-th variable deleted will be denoted by $m \setminus \{i\}$. We assume that for the considered model m matrix $(\mathbf{X}_m'\mathbf{X}_m)^{-1}$ exists.

Let $\hat{\boldsymbol{\beta}}_m = (\hat{\beta}_{i_1,m}, \ldots, \hat{\beta}_{i_{|m|},m})'$ be the least squares estimator based on model m and

$$T_{i,m} = \hat{\beta}_{i,m}[\hat{\sigma}_m^2(\mathbf{X}_m'\mathbf{X}_m)^{-1}_{i,i}]^{-1/2}, \quad i \in \{i_1, \ldots, i_{|m|}\}$$

be t-statistic corresponding to variable i when model m is fitted to the data. In the above formula $\hat{\sigma}_m^2 = (n - |m|)^{-1}\mathrm{RSS}(m)$, where $\mathrm{RSS}(m) = \mathbf{Y}'(\mathbf{I} - P_m)\mathbf{Y}$ is the sum of the squared residuals (residual sum of squares) for model m and P_m is a projection on the column space spanned by the regressors corresponding to model m. The following equality holds

$$\frac{T_{i,m}^2}{n - |m|} = \frac{\mathrm{RSS}(m \setminus \{i\}) - \mathrm{RSS}(m)}{\mathrm{RSS}(m)}. \tag{2}$$

Thus $T_{i,m}^2/(n - |m|)$ is a relative increase of RSS when variable i is dropped from the model m. It follows from (2) and generalized Cochran theorem that $T_{i,m}^2/(n - |m|)$ is a ratio of two independent chi squared distributed random variables: $\chi_1^2(\lambda_1)$ in the case of the numerator and $\chi_{n-|m|}^2(\lambda_2)$ for the denominator, where parameters of noncentrality are equal to $\lambda_1 = \|(P_m - P_{m\setminus\{i\}})\mathbf{X}\beta\|^2/(2\sigma^2)$ and $\lambda_2 = \|(I - P_m)\mathbf{X}\beta\|^2/(2\sigma^2)$, respectively.

Note also that due to a variance decomposition for a linear model which includes constant regressor we have

$$\frac{T_{i,m}^2}{n - |m|} = \frac{R_m^2 - R_{m\setminus\{i\}}^2}{1 - R_m^2}, \tag{3}$$

where R_m^2 is a coefficient of determination for a model m. Eq. (3) provides the main motivation for our choice of weights in the RSM scheme. Namely, it indicates that up to a multiplicative factor, $T_{i,m}^2$ is a decrease in R^2 due to leaving out x_i multiplied by a measure of goodness-of-fit $(1 - R_m^2)^{-1}$ of model m and thus it combines two characteristics: importance of a feature within the model m and the importance of the model itself.

In the case of random $\mathbf{X}$ the following quantities will be useful. Assume throughout for simplicity that $\mathbf{E}(x_i) = 0$ for $i \in \{1, \ldots, M\}$. Let $\mathrm{cov}(y, \mathbf{z})$ be the $1 \times |m|$ vector of covariances between y and coordinates of some $|m|$-dimensional random vector $\mathbf{z}$. Let

$$\rho_{y,\mathbf{x}_m}^2 = \frac{\mathrm{cov}^2(y, P_m y)}{\mathrm{var}(y)\mathrm{var}(P_m y)} = \frac{\mathrm{var}(P_m y)}{\mathrm{var}(y)} \tag{4}$$

be the squared multiple correlation coefficient between y and variables corresponding to model m. It is easy to see that

$$\rho_{y,\mathbf{x}_m}^2 = \frac{\mathrm{cov}(y, \mathbf{x}_m)\,\Sigma_{\mathbf{x}_m}^{-1}\mathrm{cov}(\mathbf{x}_m, y)}{\mathrm{var}(y)}, \tag{5}$$

where $\mathrm{cov}(\mathbf{x}_m, y) = \mathrm{cov}(y, \mathbf{x}_m)'$ and $\Sigma_{\mathbf{x}_m}$ is the variance–covariance matrix of variables corresponding to m. Moreover, it follows that $\rho_{y,\mathbf{x}_m}^2$ equals the maximal value of a squared correlation between y and linear combination of $\mathbf{x}_m$, when the

coefficients of the combination vary. For $m = \{i\}$ consisting of one element $\rho^2_{y,\mathbf{x}_m}$ is squared correlation coefficient $\rho^2(y, x_i)$ between variables y and x_i.

In the case of deterministic $\mathbf{X}$ let $\lambda_m := \lim_{n\to\infty} n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_m\mathbf{X}\boldsymbol{\beta}\|^2$. Note that $\|\mathbf{X}\boldsymbol{\beta} - P_m\mathbf{X}\boldsymbol{\beta}\|^2$ equals a squared distance of $\mathbf{X}\boldsymbol{\beta}$ from its projection $P_m\mathbf{X}\boldsymbol{\beta}$ on the columns of $\mathbf{X}$ corresponding to m and may be regarded as a measure of discrepancy between the larger and the smaller model. Thus λ_m is a limiting average value of this discrepancy per one coordinate of $n \times 1$ vector $\mathbf{X}\boldsymbol{\beta}$. Remark 1 gives another interpretation of λ_m in terms of a limiting prediction error.

The following theorem shows that ordering variables with respect to squares of their t-statistics is in the case of deterministic $\mathbf{X}$ asymptotically equivalent to ordering with respect to quantities $\lambda_{m\setminus\{i\}}$. It also turns out that in the case of random $\mathbf{X}$ under appropriate moment conditions $\lambda_{m\setminus\{i\}}$ exist almost surely and the ordering can be reexpressed in terms of squared multiple correlation coefficients $\rho^2{}_{y,\mathbf{x}_{m\setminus\{i\}}}$. In the following number of fitted variables m is a fixed integer. Note that as $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_t\boldsymbol{\beta}_t$, λ_m does not depend on the number of potential regressors M. The same observation applies to $T^2_{i,m}$.

Theorem 1. *Let $i, j \in m$.*

(i) *In the case of deterministic $\mathbf{X}$ assume that $\lambda_{m\setminus\{i\}}$ and $\lambda_{m\setminus\{j\}}$ exist. Then $T^2_{i,m} > T^2_{j,m}$ almost surely for sufficiently large n iff*

$$\lambda_{m\setminus\{i\}} > \lambda_{m\setminus\{j\}}. \tag{6}$$

(ii) *In the case of random $\mathbf{X}$ assume that $\Sigma_{\mathbf{x}_m}$ is invertible and $\mathbf{E}x_j^4$ are finite for all $j \in m$. Then $T^2_{i,m} > T^2_{j,m}$ almost surely for sufficiently large n iff*

$$\rho^2_{y,\mathbf{x}_{m\setminus\{j\}}} > \rho^2_{y,\mathbf{x}_{m\setminus\{i\}}}. \tag{7}$$

In the case of random $\mathbf{X}$ the explicit formula for almost sure limits in (6) can be obtained and condition (6) is simplified to (7) (see the proof of Theorem 1). It is also easy to see that for m having two elements condition (7) is equivalent to $\rho^2(y, x_i) = \rho^2(y, \mathbf{x}_{m\setminus\{j\}}) > \rho^2(y, \mathbf{x}_{m\setminus\{i\}}) = \rho^2(y, x_j)$.

Remark 1. When $\mathbf{X}$ is deterministic consider the mean squared error of prediction for model m

$$\mathrm{MSEP}_n(m) = \mathbf{E}(\|\mathbf{Y}^* - \mathbf{X}_m\hat{\boldsymbol{\beta}}_m\|^2) = \sigma^2(n + |m|) + \|\mathbf{X}\boldsymbol{\beta} - P_m\mathbf{X}\boldsymbol{\beta}\|^2,$$

where $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$ with $\boldsymbol{\varepsilon}^*$ being an independent copy of $\boldsymbol{\varepsilon}$. Let $\mathrm{MSEP}(m) = \lim_{n\to\infty} n^{-1}\mathrm{MSEP}_n(m)$. Thus the ordering in (6) is equivalent to ordering

$$\mathrm{MSEP}(m \setminus \{i\}) > \mathrm{MSEP}(m \setminus \{j\}).$$

Moreover, for random $\mathbf{X}$, due to (4) and the last equality in (12), (7) is equivalent to

$$\mathrm{var}(y - P_{m\setminus\{i\}}y) > \mathrm{var}(y - P_{m\setminus\{j\}}y).$$

Remark 2. In the case of deterministic $\mathbf{X}$ let

$$t_{i,m} = \frac{\lambda_{m\setminus\{i\}} - \lambda_m}{\sigma^2 + \lambda_m} = \frac{\mathrm{MSEP}(m \setminus \{i\}) - \mathrm{MSEP}(m)}{\mathrm{MSEP}(m)}$$

and for the random $\mathbf{X}$

$$t_{i,m} = \frac{\rho^2_{y,\mathbf{x}_m} - \rho^2_{y,\mathbf{x}_{m\setminus\{i\}}}}{1 - \rho^2_{y,\mathbf{x}_m}}.$$

It follows from the proof of Theorem 1 that under its assumptions in both cases $(n - |m|)^{-1}T^2_{i,m} \xrightarrow{\text{a.s.}} t_{i,m}$.

Corollary 1. *Let $m \supseteq t$.*

(i) *In the case of deterministic $\mathbf{X}$ assume that $\lambda_{m\setminus\{i\}}$ is defined for any i. Then $\min_{i\in t} T^2_{i,m} > \max_{i\in t^c\cap m} T^2_{i,m}$ almost surely for sufficiently large n iff*

$$\lambda_{m\setminus\{i\}} > 0, \tag{8}$$

for all $i \in t$.

(ii) *In the case of random $\mathbf{X}$ assume that $\Sigma_{\mathbf{x}_m}$ is invertible and $\mathbf{E}x_j^4 < \infty$ for all $j \in m$. Then $\min_{i\in t} T^2_{i,m} > \max_{i\in t^c\cap m} T^2_{i,m}$ almost surely for sufficiently large n.*

Various versions of condition (8) are used to prove asymptotic results of model selection for linear models (cf. Zhang, 1992; Shao, 1993; Zheng and Loh, 1995; Casella et al., 2009). e.g. in the last paper the condition equivalent to $\lambda_s > 0$ for any s such that $t \not\subset s$ is used to prove consistency of the Bayes selection method introduced there. Here we use a more general

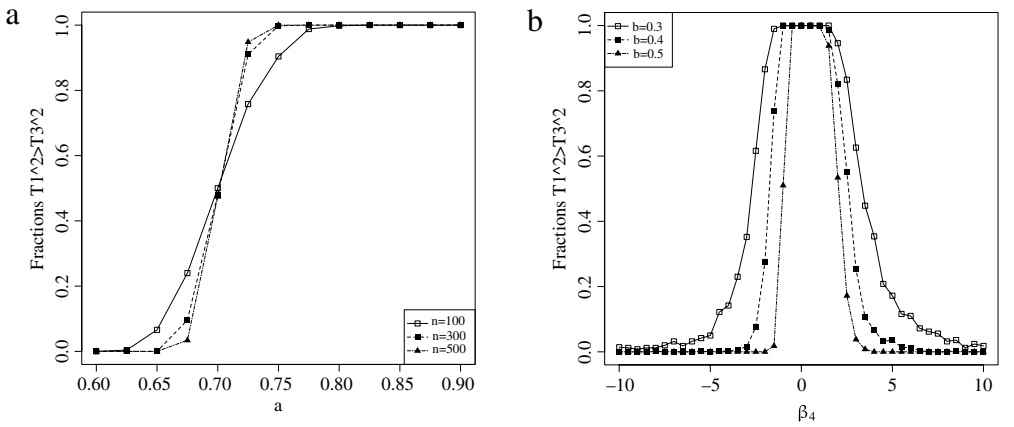

Fig. 1. (a) Example 1: estimated probabilities of $T^2_{2,m} > T^2_{3,m}$ with respect to a based on $N = 500$ trials. (b) Example 2: estimated probabilities of $T^2_{1,m} > T^2_{3,m}$ with respect to β_4 based on $N = 500$ trials.

condition (6) in order to show that the ordering induced by the squared t-values is asymptotically equivalent to the ordering by λ values. Note the fact that (8) is automatically satisfied for random $\mathbf{X}$ which can be regarded of superior feature of random design when compared to fixed design modelling.

Corollary 2. *Assume that $\Sigma_{\mathbf{x}_{m\cup t}}$ is diagonal, invertible and $\mathbf{E}x_j^4 < \infty$ for all $j \in m$ (in the case of random $\mathbf{X}$) and $\lim_{n\to\infty} n^{-1}\mathbf{X}'_{m\cup t}\mathbf{X}_{m\cup t}$ is diagonal and invertible (in the case of deterministic $\mathbf{X}$). Then $\min_{i\in t\cap m} T^2_{i,m} > \max_{i\in t^c\cap m} T^2_{i,m}$.*

Corollaries 1 and 2 indicate that, when a model containing all significant variables is fitted or variables are uncorrelated, the ordering with respect to the squared t-statistics ensures that the coordinates corresponding to nonzero coefficients are placed ahead the spurious ones. In general case when the fitted model is misspecified (i.e. at least one significant variable is omitted) and the variables are not independent it may happen that condition (6) or (7) is not satisfied for some $i \in t, j \notin t$ and irrelevant variable j is placed ahead relevant variable i when the ordering of variables is based on squared t-statistics. Examples 1 and 2 explore such situations for two different dependence structures of attributes.

Example 1. Consider random-design regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, \beta_4)'$, $\boldsymbol{\varepsilon}$ has $N(0, \mathbf{I})$ distribution and rows of $\mathbf{X}$ are normally distributed with covariance matrix

$$\Sigma_{\mathbf{x}} = (\sigma_{ij}) = \begin{bmatrix} 1 & a & 0.8 & a \\ a & 1 & 0.8 & a \\ 0.8 & 0.8 & 1 & 0.8 \\ a & a & 0.8 & 1 \end{bmatrix},$$

where $a \in (0, 1)$ is a parameter. Thus all relevant variables are equicorrelated with correlation equal a and their correlation with spurious variable x_3 equals 0.8. A misspecified model $y \sim x_2 + x_3$ containing two variables only is fitted: x_2 (relevant) and x_3 (spurious). Theorem 1(ii) states that $T^2_{2,m} > T^2_{3,m}$ for sufficiently large n with probability 1 i.e. the relevant variable will precede the spurious one in the ordering if and only if (7) is satisfied. It is easy to verify that in this case condition (7) yields

$$\sigma_{22}^{-1}[\beta_1\sigma_{12} + \beta_2\sigma_{22} + \beta_4\sigma_{24}]^2 > \sigma_{33}^{-1}[\beta_1\sigma_{13} + \beta_2\sigma_{23} + \beta_4\sigma_{34}]^2,$$

or equivalently $\rho^2(x_2, y) > \rho^2(x_3, y)$. For $\beta_1 = \beta_2 = \beta_4 = 1$ an easy calculation shows this result in $a > 0.7$. The intuitive reason is that when relevant variables x_1 and x_4 missing from the model become less correlated their individual contributions are more significant. As spurious variable x_3 is strongly correlated with both of them it takes over their roles in the misspecified model and in effect has more predictive power than variable x_2. For $\beta_1 = \beta_2 = \beta_4 = 1$ we carried out $L = 500$ simulations for $n = 100, 200, 500$ and computed fraction of correct orderings for which $T^2_{2,m} > T^2_{3,m}$ with changing value of parameter a. The results are presented in Fig. 1(a). When the correlation between spurious variable x_3 and relevant variables x_1 and x_4 missing from the model is strong and the correlation between variable x_1 and x_4 is relatively weak then ordering of variables in m induced by t-statistics can be incorrect with high probability, i.e. it is likely that $T^2_{2,m} < T^2_{3,m}$.

Example 2. Consider random-design regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, 0, \beta_4)'$, $\boldsymbol{\varepsilon}$ has $N(0, \mathbf{I})$ distribution and rows of $\mathbf{X}$ are normally distributed with covariance matrix

$$\Sigma_{\mathbf{x}} = \begin{bmatrix} 1 & b & 0 & 0 \\ b & 1 & b & 0 \\ 0 & b & 1 & b \\ 0 & 0 & b & 1 \end{bmatrix},$$

where $b \in (0, 0.5]$ is a parameter. A misspecified model $y \sim x_1 + x_2 + x_3$ containing three variables is fitted: x_1, x_2 (relevant) and x_3 (spurious). We set $\beta_1 = \beta_2 = 1$ and β_4 is treated as a parameter. In this case only adjacent predictors are correlated with the correlation coefficient equal to b. In this case condition (7) for variables x_1 and x_3 takes the form

$$(1+b)^2 - 2b(1+b)^2 > (b+\beta_4 b)^2 - 2b(1+b)(b+\beta_4 b).$$

The above inequality holds for $\beta_4 \in (\frac{2b^2-1}{b}, \frac{1}{b})$. We carried out $L = 500$ simulations for $b = 0.3, 0.4, 0.5$ and computed fractions of simulations in which $T^2_{1,m} > T^2_{3,m}$ as a function of β_4. Large values of parameter β_4 corresponding to omitted variable result in increased probability of incorrect ordering. Moreover, for a fixed β_4 when correlation b between x_3 and x_4 increases the probability of correct ordering decreases. The results are presented in Fig. 1(b).

3. Random subspace method

We first describe the algorithm of the random subspace method with the weighting pertaining to values of squared t-statistics. It follows from the description that a parallel version of the algorithm is very easy to implement.

Algorithm. 1. Input: observed data $(\mathbf{Y}, \mathbf{X})$, number of subset draws B, size of the subspace $|m| < M$.
2. Repeat the following procedure for $k = 1, \ldots, B = B_n$, where B_n is such that $B_n \to \infty$ when $n \to \infty$ and starting with $C_{i,0} = 0$ for any i.
   - Randomly draw a model $m^* = \{i^*_1, \ldots, i^*_{|m|}\}$ from the original feature space.
   - Fit model $y \sim \mathbf{x}_{m^*}$ and compute T^2_{i,m^*} for each $i \in m^*$. Set $T^2_{i,m^*} = 0$ if $i \notin m^*$.
   - Update the counter $C_{i,k} = C_{i,k-1} + I\{i \in m^*\}$.
3. For each variable i compute the final t-based score TS^*_i defined as

$$\mathrm{TS}^*_i = \frac{1}{C_{i,B}} \sum_{m^*: i \in m^*} \frac{T^2_{i,m^*}}{n - |m|}.$$

4. Sort the list of variables according to scores TS^*_i : $\mathrm{TS}^*_{i_1} \geq \mathrm{TS}^*_{i_2} \ldots \geq \mathrm{TS}^*_{i_M}$.
5. Output: Ordered list of variables $\{i_1, \ldots, i_M\}$.

Let $\mathcal{M}_{|m|}$ be the family of all subsets $\{i_1, \ldots, i_{|m|}\}$ of $\{1, \ldots, M\}$ (models) of size $|m|$ and $|\mathcal{M}_{|m|}| = \binom{M}{|m|}$ be its cardinality. Analogously let $\mathcal{M}_{i,|m|}$ be the family of all subsets (models) of size $|m|$ containing variable i and $|\mathcal{M}_{i,|m|}| = \binom{M-1}{|m|-1}$. We define resampling measure P^* on $\mathcal{M}_{|m|}$ such that for any model $m \in \mathcal{M}_{|m|}$ we have $P^*(m) = \frac{1}{|\mathcal{M}_{|m|}|}$. The expected value with respect to this distribution will be denoted by $\mathbf{E}^*$. We state first the result for the case when the number of predictors M is fixed.

Theorem 2. *In the case of deterministic* $\mathbf{X}$ *assume that* λ_m *and* $\lambda_{m\setminus\{i\}}$, $i \in m$, *exist for all subsets of a given size* $|m|$. *In the case of random* $\mathbf{X}$ *assume that* $\Sigma_{\mathbf{x}_m}$ *is invertible for all subsets of a given size* $|m|$ *and* $\mathbf{E}x_j^4 < \infty$ *for all* j. *Then for almost any sequence* $(\mathbf{Y}_n, \mathbf{X}_n)_{n=1}^{\infty}$

$$\mathrm{TS}^*_i \xrightarrow{P^*} ts_i := \frac{1}{|\mathcal{M}_{i,|m|}|} \sum_{m \in \mathcal{M}_{i,|m|}} t_{i,m}, \quad \text{as } n \to \infty,$$

where $t_{i,m}$ *is defined in Remark 2.*

Thus TS^*_i is asymptotically equivalent to weight ts_i of variable i equal to a relative increment of mean squared error of prediction MSEP, when the variable is omitted from model m, averaged over all models m containing it.

Corollary 3. *Assume that* $\Sigma_{\mathbf{x}}$ *is invertible and diagonal and* $\mathbf{E}x_j^4 < \infty$ *for all* j *(in the case of random* $\mathbf{X}$*) and* $\lim_{n\to\infty} n^{-1}\mathbf{X}'\mathbf{X} = W$, *where* W *is invertible and diagonal (in the case of deterministic* $\mathbf{X}$*). Then for almost any sequence* $(\mathbf{Y}_n, \mathbf{X}_n)_{n=1}^{\infty}$

$$P^*(\min_{i \in t} \mathrm{TS}^*_i > \max_{i \notin t} \mathrm{TS}^*_i) \to 1 \quad \text{as } n \to \infty.$$

Remark 3. When the first column of design matrix $\mathbf{X}$ consists of ones and corresponds to an intercept in the model, a slight modification of the random subspace method is necessary. When randomly sampling a subset m we sample from genuine regressors only and then a model pertaining to regressors from m and an intercept is fitted. Theorem 2 and Corollary 3 remain valid also in this case.

We now state the result for the case when the number of predictors M_n may depend on n under more stringent assumptions on design.

Theorem 3. *In the case of deterministic* $\mathbf{X}$ *assume that* (i) $\max_{s\subset\{1,\ldots,M_n\}} |n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2 - \lambda_s| \to 0$, *when* s *is a subset of the fixed size* $|m|$ *or* $|m|-1$, (ii) $\log(M_n) = o(n)$ *and* (iii)

$$\limsup_n \frac{1}{|\mathcal{M}_{i,|m|}|} \sum_{m\in\mathcal{M}_{i,|m|}} t_{i,m} < \infty.$$

In the case of random $\mathbf{X}$ *assume that* (i′) *the minimal eigenvalue of* $\Sigma_{\mathbf{x}_s}$ *is bounded away from* 0 *for all subsets* s *of the fixed size* $|m|$ *or* $|m|-1$, (ii′) $M_n = O(n^\alpha)$ *where* α *is such that* $\max_{j\le M_n} \mathbf{E}|x_j|^{4\alpha+4+\delta} \le C$ *for some* $\delta > 0$ *and* (iii). *Then*

$$\mathrm{TS}_i^* - ts_i \xrightarrow{P^*} 0 \quad \text{as } n\to\infty,$$

where ts_i *is defined in Theorem 2.*

Remark 4. For deterministic design the result holds for M_n satisfying very mild condition $M_n = o(\exp(n))$. It follows from the proof of Theorem 3 that in the case of random $\mathbf{X}$ conditions (i′) and (ii′) imply that (i) is satisfied a.s. In this case we can also assume milder condition (ii) on the growth of M_n instead of $M_n = O(n^\alpha)$ at the expense of assuming that variables x_j^2 satisfy Cramér condition uniformly for $j = 1,\ldots,M_n$.

We now state the analogue of Corollary 3 for $M = M_n$.

Corollary 4. *Assume that conditions of Theorem 3 are satisfied,* $\Sigma_{\mathbf{x}_s}$ *is invertible and diagonal and* $\mathbf{E}x_j^4 < \infty$ *for all* j *(in the case of random* $\mathbf{X}$*) and* $\lim_{n\to\infty} n^{-1}\mathbf{X}_s'\mathbf{X}_s = W_s$, *where* W_s *is invertible and diagonal (in the case of deterministic* $\mathbf{X}$*), and* s *is any finite set. Then for almost any sequence* $(\mathbf{Y}_n, \mathbf{X}_n)_{n=1}^\infty$ *and any* $k\in\mathbf{N}$

$$P^*(\min_{i\in t} \mathrm{TS}_i^* > \max_{i:\, i\notin t\wedge i\le k} \mathrm{TS}_i^*) \to 1 \quad \text{as } n\to\infty.$$

Corollary 4 asserts that when variables are asymptotically uncorrelated we can order with probability tending to 1 any *finite* sequence of variables in such a way that true variables will precede spurious ones. Note however that when $M > n$ it is in general not possible to select true variables among all possible regressors consistently (cf. Shao and Deng, 2012 and Fan and Lv, 2008).

4. Practical performance of the proposed method

4.1. Model selection and prediction

We briefly describe model selection procedure based on the RSM, the lasso method and the univariate approach. In the following observed data $(\mathbf{Y}, \mathbf{X})$ is split into two subsets: training set $(\mathbf{Y}^t, \mathbf{X}^t)$ containing n_t observations and validation set $(\mathbf{Y}^v, \mathbf{X}^v)$ containing n_v observations. Let also $(\mathbf{Y}^{\mathrm{new}}, \mathbf{X}^{\mathrm{new}})$ containing n_{new} observations be a test set.

Random subspace method

The algorithm described in Section 3 is performed on training data $(\mathbf{Y}^t, \mathbf{X}^t)$ and the covariates indexed by $\{i_1,\ldots,i_M\}$ are ordered with respect to decreasing values of the scores

$$\mathrm{TS}_{i_1}^* \ge \mathrm{TS}_{i_2}^* \ge \cdots \ge \mathrm{TS}_{i_M}^*.$$

From the hierarchical list of models $\{\{i_1\},\{i_1,i_2\},\ldots,\{i_1,\ldots,i_{\min(n_t,M)}\}\}$ we select model $m_{\mathrm{opt}} = \{i_1,\ldots,i_{|m_{\mathrm{opt}}|}\}$ for which the prediction error $n_v^{-1}\|\mathbf{Y}^v - \mathbf{X}^v\hat{\boldsymbol{\beta}}_{m_{\mathrm{opt}}}\|^2$ is minimal. Here, $\hat{\boldsymbol{\beta}}_{m_{\mathrm{opt}}}$ is a least squares estimator based on model m_{opt} computed on training data. Two parameters need to be set in the RSM: the number of selections B and the subspace size $|m|$. The smaller the size of a chosen subspace (i.e. a subset of features chosen) the larger the chance of missing informative features or missing dependences between variables. On the other hand, for large $|m|$ many spurious variables can be included adding noisy dimensions to the subspace. Note that the subspace size is also limited by the number of observations, namely must be not larger than n_t. Here the value of parameter $|m|$ is chosen empirically. We concluded from numerical experiments that the reasonable choice is $|m| = \min(n_t, M)/2$. This is also confirmed by real data examples (see Figs. 6(b) and 7(c)). The performance of prediction error with respect to $|m|$ presented in Figs. 6(b) and 7(c) is typical also for artificial datasets discussed in Section 4.3.

Lasso method

The lasso estimate is defined by (cf. Tibshirani, 1996)

$$\hat{\boldsymbol{\beta}}_{\mathrm{lasso}}(\alpha) = \arg\min_{\boldsymbol{\beta}} \left[\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \alpha\|\boldsymbol{\beta}\|_{l_1}\right],$$

where $\|\cdot\|_{l_1}$ denotes l_1 norm and α is a parameter. Because of the nature of the penalty choosing sufficiently large α will result in some of the coefficients to be exactly zero. Thus the lasso can be viewed as a variable selection method. The optimal value α (denoted by α_{opt}) is chosen by minimizing the prediction error on independent validation set, i.e. $n_v^{-1}\|\mathbf{Y}^v - \mathbf{X}^v\hat{\boldsymbol{\beta}}_{\mathrm{lasso}}(\alpha)\|^2$

or by cross-validation. We choose the first option in our numerical experiments in order to make a comparison with the RSM more objective.

Univariate method

The univariate approach is considered here as a benchmark. In this method informativeness and prediction strength of each feature is evaluated individually. Here, for each variable $i \in \{1, \ldots, M\}$ we compute squared value of its t-statistic $T^2_{i,\{i\}}$ based on simple regression model $y \sim x_i$. Then the covariates are ordered with respect to $T^2_{i,\{i\}}$ and the same procedure on hierarchical list of models as in the RSM is performed.

For all methods described above the prediction strength of the selected model is assessed by prediction error on independent test set using the average error $n^{-1}_{\text{new}}\|\mathbf{Y}^{\text{new}} - \mathbf{X}^{\text{new}}\hat{\boldsymbol{\beta}}_{m_{\text{opt}}}\|^2$ with $\hat{\boldsymbol{\beta}}_{m_{\text{opt}}}$ being an estimator based on model m_{opt} computed on training data.

4.2. Variable importance estimation

We start by describing two methods of estimating variable importance which are compared with RSM for which importance of variable x_i is estimated as TS^*_i.

Measures based on the random forests

For a given regression tree t 'out of bag' MSE_t for this tree is calculated as an average of squared differences between predictions and actual values of observations which have not taken part in construction of t. For a given variable x_i $\text{MSE}_{t,i}$ is an analogous quantity with the difference that the tree t is now constructed using data with coordinate corresponding to x_i randomly permuted. Weight WB1_i equals an average of $\text{MSE}_t - \text{MSE}_{t,i}$ over all trees. The second weight WB2_i proposed by Breiman pertains to an average decrease of RSS for all knots for which x_i was a splitting variable. For details we refer to Breiman (2001), see however also Sandri and Zuccolotto (2010), who showed that measure WB2_i can be affected by bias due to the lack of pruning for trees used in the random forests.

Measure based on MARS

MARS model, being an adaptation of the CART method to improve its performance in regression setting, is fitted to data. It consists in a stepwise procedure which at each step constructs a new larger function basis such that new elements are products of elements of the prior basis and linear splines with knots at coordinates of data points. The resulting large fitted model is then pruned using Generalized Cross-validation (GCV) criterion (for a detailed description see e.g., Hastie et al., 2009). Variable importance is defined as cumulative value of a change of GCV calculated over steps when the term involving the variable in question is added to the model. In numerical experiments R package `caret` has been used to calculate it (cf. Kuhn, 2008).

We also shortly discuss `lmg` measure as it bears some similarity with our approach. Assume that the number of potential attributes M is smaller than n and let $r \in \Sigma(M)$ be an arbitrary permutation of $\{1, 2, \ldots, M\}$. Moreover denote by $s_i(r)$ the set of indices of variables which precede variable x_i in the ordering r i.e. $s_i(r) = \{j : r(j) < r(i)\}$ and let $\Delta R^2_{i,r} = R^2_{s_i(r)\cup\{i\}} - R^2_{s_i(r)}$. Thus $\Delta R^2_{i,r}$ is a sequential increase of R^2 corresponding to adding variable x_i in the ordering induced by permutation r. Weight assigned to x_i is $\text{lmg}_i = M!^{-1}\sum_{r\in\Sigma(M)} \Delta R^2_{i,r}$. Note two important differences with our proposal. First, by considering $\Delta R^2_{i,r}$ only the importance of x_i in the model is taken into account without evaluating the goodness of fit of the model. Moreover, in the above definition the weight is averaged over all permutations whereas in the RSM it is averaged over all models of cardinality $|m|$. Actually, it is easy to see that averaging over all models of cardinality $|m|$ is equivalent to averaging over all permutations such that $r(i) = |m|$. `lmg` measure was not considered in our numerical experiments as it is computationally too expensive for number of predictors considered (M equal to 100 and 1000).

4.3. Numerical experiments

In this section, the performance of the discussed methods, described in Section 4.1, is investigated. Our main objective is to study the performance of the proposed method as a prediction approach. Random forests are also used as a benchmark method for prediction and weight assignment in their original form which does not involve variable selection. Moreover the weight assignment by the RSM is compared with other approaches. We considered moderate sample size $n = 200$.

The artificial datasets

Recall that t denotes the set of coordinates which correspond to non-zero coefficients $\boldsymbol{\beta}_t$. The following linear models have been considered:

(M1) $t = (1, 5, 10)$, $\boldsymbol{\beta}_t = (1, 1, 1)'$,
(M2) $t = (1, 5, 10, 15, 20, 25, 30)$, $\boldsymbol{\beta}_t = (2, 2, 2, 2, -2, -2, -2)'$,
(M3) $t = (1, 5, 10, 15, 20, 25, 30, 35, 40, 45)$, $\boldsymbol{\beta}_t = (3, 3, 3, 3, 3, -3, -3, -3, -3, -3)'$,
(M4) $t = (1, \ldots, 5, 11, \ldots, 15, 21, \ldots, 25)$, $\boldsymbol{\beta}_t = (2.5, 2.5, 2.5, 2.5, 2.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1, 1, 1, 1, 1)'$.

Number of potential regressors is either $M = 100$ or 1000, e.g. in the first case for model M1 there is 97% of redundant regressors. The rows of $\mathbf{X}$ were generated independently from the standard normal M-dimensional distribution with zero mean and the covariance matrix $\Sigma_{\mathbf{x}} = (\rho_{ij}) = \rho^{|i-j|}$. The outcome is $\mathbf{Y} = \mathbf{X}_t\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ has zero-mean normal

Table 1
Positive selection rates and false discovery rates based on $L = 200$ simulation trials. Table (a) on the left-hand side corresponds to $\rho = 0$ whereas the table (b) on the right-hand side to $\rho = 0.8$.

(a)

		M1	M2	M3	M4
RSM	PSR	1	1	1	0.97
	FDR	0.06	0.13	0.10	0.42
UNIVARIATE	PSR	1	1	1	0.91
	FDR	0.15	0.31	0.51	0.65
LASSO	PSR	1	1	1	1
	FDR	0.76	0.71	0.69	0.64

(b)

		M1	M2	M3	M4
RSM	PSR	1	1	1	0.958
	FDR	0.25	0.16	0.16	0.13
UNIVARIATE	PSR	1	0.99	0.99	0.97
	FDR	0.62	0.77	0.79	0.44
LASSO	PSR	1	1	1	0.99
	FDR	0.71	0.69	0.66	0.41

distribution with covariance matrix $\sigma^2\mathbf{I}$ and $\sigma^2 = 1$ (for models M1–M3) and $\sigma^2 = 1.5$ (for model M4). Our model M4 is analogous to model 7 in Huang et al. (2008). Observe that for models M1–M3 when $\rho > 0$ dependence between the relevant variables is much weaker than that between the relevant variables and the spurious ones adjacent to them. The simulation experiments were repeated $L = 200$ times. For each simulation trial data $(\mathbf{Y}, \mathbf{X})$ is split into training set $(\mathbf{Y}^t, \mathbf{X}^t)$ and validation set $(\mathbf{Y}^v, \mathbf{X}^v)$ containing $n/2 = 100$ observations each and final model m_{opt} is selected as described in Section 4.1. For the RSM we considered $B = 1000$ choices of a random subspace consisting of $|m| = \min(n_t, M)/2 = 50$ attributes.

As the measures of performance, besides the prediction error on independent test data containing 100 observations, we also consider the positive selection rate (PSR) defined as $E(|m_{\text{opt}} \cap t|/|t|)$ and the false discovery rate (FDR) $E(|m_{\text{opt}} \setminus t|/|m_{\text{opt}}|)$. Note that FDR measures a fraction of false positives with respect to all positives. We first discuss the case $M = 100$. The results for $\rho = 0$ and $\rho = 0.8$, presented in Fig. 2, are encouraging and indicate that the RSM works better than the lasso and the univariate method when dependence between predictors is strong. In the case of independent covariates the RSM also outperforms other methods with the exception of model M4 for which lasso works slightly better. The results of t-test with $\alpha = 0.05$ indicate that means of prediction errors for RSM are significantly smaller than the corresponding means for lasso in all considered cases except model M4 with $\rho = 0$. For models M2–M4 the univariate method works uniformly worst in the case $\rho = 0.8$. For models M2–M4 with $\rho = 0.8$ application of the univariate method resulted in about 5% outliers, i.e. unusually large prediction errors, which are not shown in Fig. 2. The results for $\rho = 0.5$ (not presented here) are similar to those for $\rho = 0.8$ but less pronounced. We have not included the results for random forests discussed later for real datasets as they are not specifically designed for linear models and perform poorly when compared with the lasso and the RSM.

We also repeated all experiments for significantly larger number of potential regressors, namely $M = 1000$. Fig. 3 shows the results for models M1 and M4, the boxplots for other cases are omitted to save space. The results in all cases M1–M4 are similar to those for $M = 100$ except for model M4 with $\rho = 0$, where lasso works now significantly better than RSM. Differences in means are again statistically significant. Observe that for all methods and models PSR is about 1 (see the results in Table 1) which indicates that almost all cases a model containing practically all variables from t is chosen.

On the other hand, the values of FDR for the RSM indicate that the number of spurious variables in the final model is significantly smaller in the case of the RSM than for the lasso and the univariate method. This seems to be a very promising feature of the RSM, especially in cases when screening of non-significant variables is costly. Note that one would expect that a smaller prediction error is associated with a larger number of false positives. Fig. 4 indicates that it is not always the case. It shows prediction errors averaged over 200 runs for model M1 with $\rho = 0$, $M = 1000$ and $n = 200$ against the number of variables included in the model, when variables have been first ordered by the RSM or the lasso. In the case of the RSM prediction curve has a clear minimum at the true number of significant variables equal to 3, whereas the corresponding minimum for the lasso is attained at much larger number of variables equal to 23.

Weight assignment methods described in Section 4.2 are discussed here for model M1 with $M = 1000$ and $n = 200$, the performance for other models was similar. Fig. 5 presents the normalized weights (i.e. divided by their sum) averaged over $L = 200$ simulations for $\rho = 0$ and $\rho = 0.8$. It is seen that in the case of independent features the significant variables (1, 5, and 10) are most apparent for MARS and WB2 measures, although the distinction between relevant and spurious variables is excellent for all weight assignment methods. When the dependence is strong ($\rho = 0.8$) the weights corresponding to spurious variables become relatively larger. Observe that when $\rho = 0.8$ the differences between weights corresponding to relevant and spurious covariates are most significant in the case of the MARS-based measure and RSM. In the case of $\rho = 0$ fractions of correct orderings are: 1 (RSM), 0.165 (WB1), 0.78 (WB2) and 1 (MARS); in the case of $\rho = 0.8$: 0.56 (RSM), 0.12 (WB1), 0.095 (WB2) and 0.94 (MARS). All differences between means corresponding to RSM, MARS and WB1 or WB2 are significant, except the means for MARS and RSM when $\rho = 0$.

Real data example I

The algorithms are also compared on real dataset Ozone which describes the relationship between atmospheric ozone concentration and meteorological indicators in the Los Angeles basin. The data consist of daily measurements of ozone concentration (maximum one hour average) and 9 meteorological quantities for $n = 330$ days of 1976. For detailed description of predictors, see Breiman and Friedman (1985). Since the relations between variables are nonlinear we fit model with interaction terms $x_i x_j$ and quadratic terms x_i^2 which together with an intercept yields 55 independent variables. We

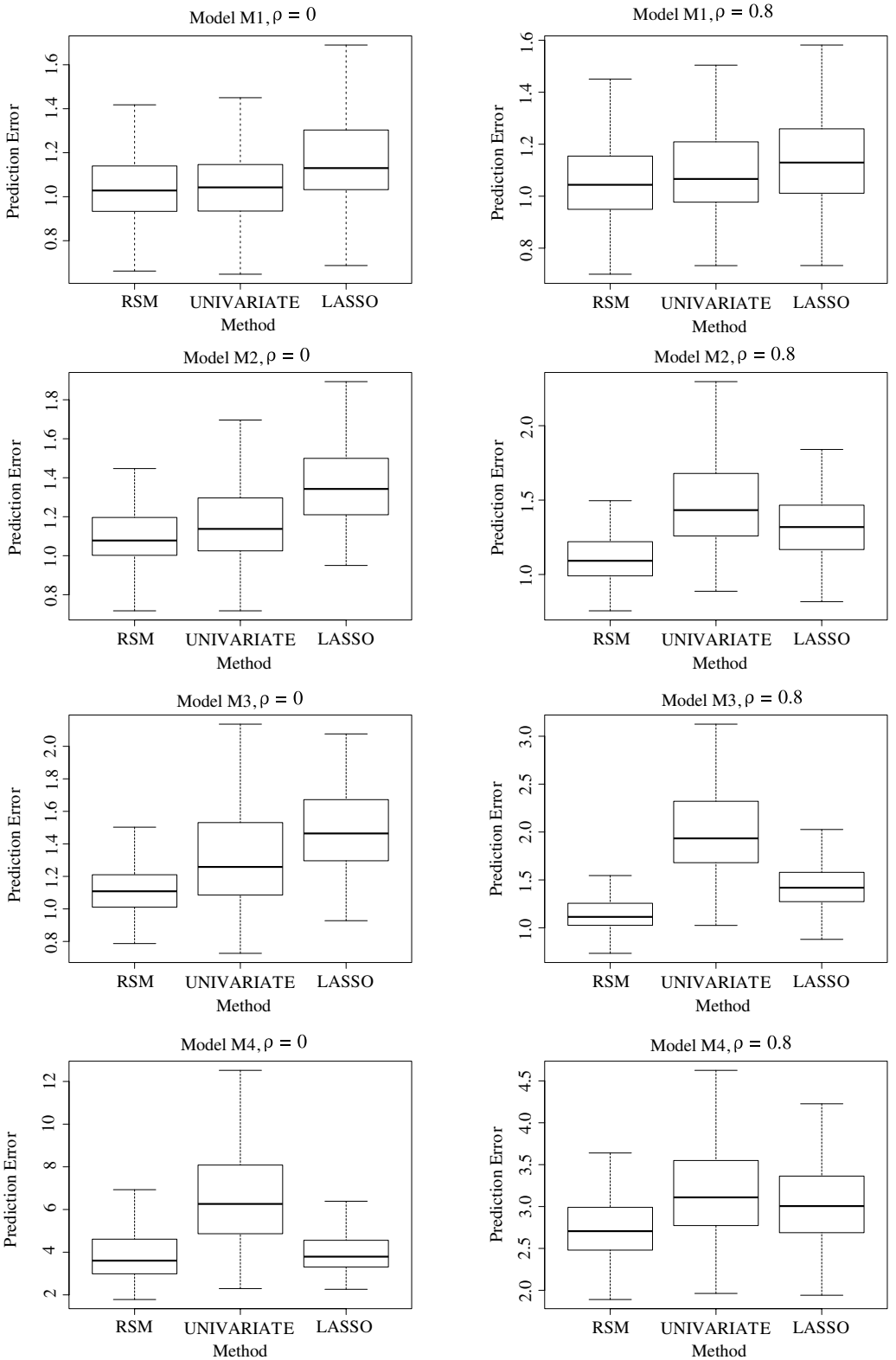

Fig. 2. Prediction errors for models M1, M2, M3, M4 with $M = 100$ and $n = 200$ based on $L = 200$ simulation trials. Left-hand side figures correspond to $\rho = 0$ whereas right-hand side to $\rho = 0.8$.

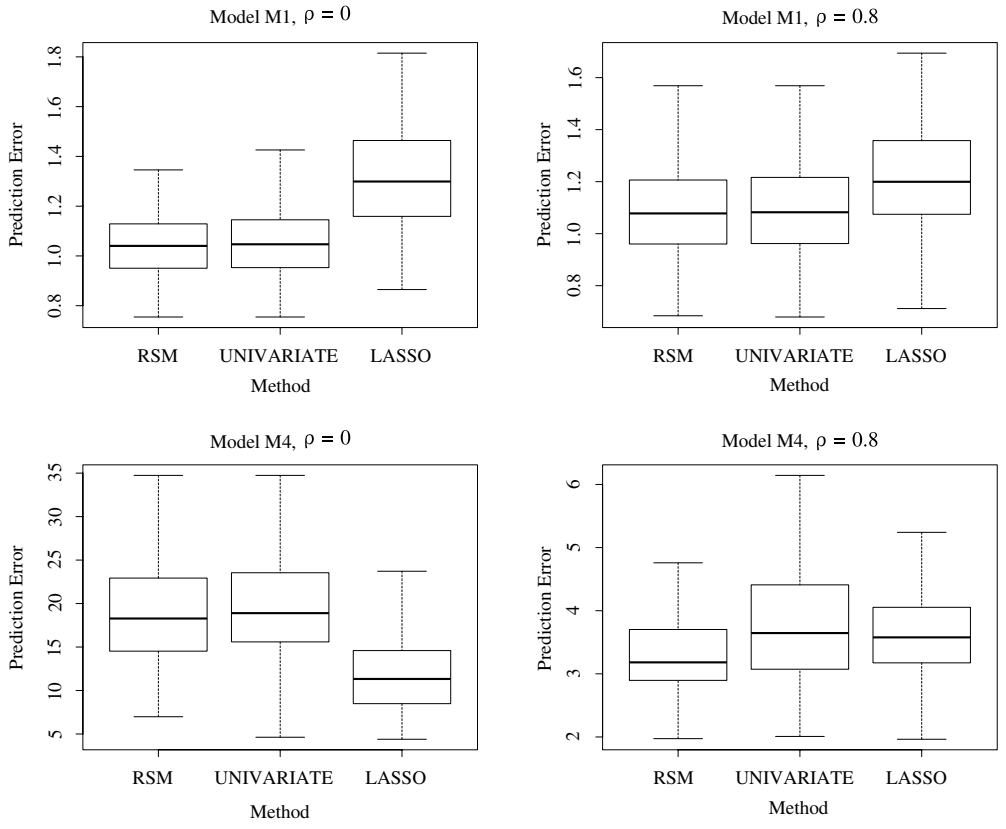

Fig. 3. Prediction errors for models M1, M4 with $M = 1000$ and $n = 200$ based on $L = 200$ simulation trials. Left-hand side figures correspond to $\rho = 0$ whereas right-hand side to $\rho = 0.8$.

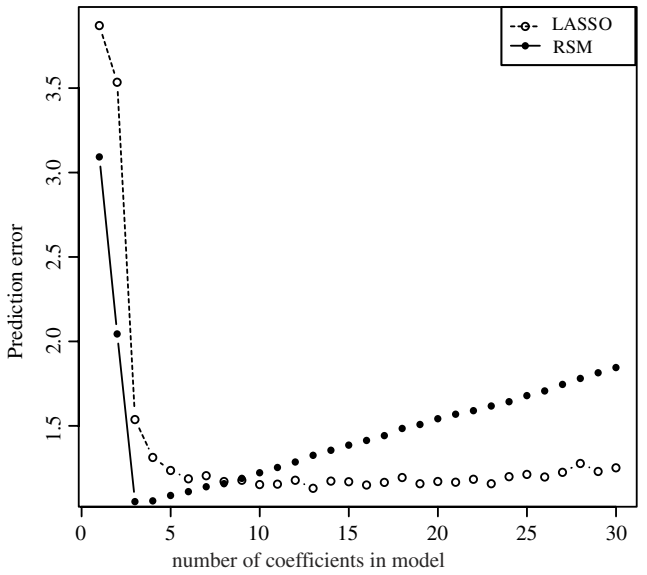

Fig. 4. Prediction errors computed on validation set for model M1 with $M = 1000, n = 200$ ($n_t = 100, n_v = 100$) with respect to the number of coefficients in the model. The errors are averaged over $L = 200$ simulation trials.

randomly split the dataset into training set of size 100, validation set of size 100 and test set of size 130. The considered methods are performed on the training set then an optimal model is selected using the validation set and finally prediction error is computed on the test set. The above procedure is repeated 200 times. For the RSM we took $B = 1000$ and

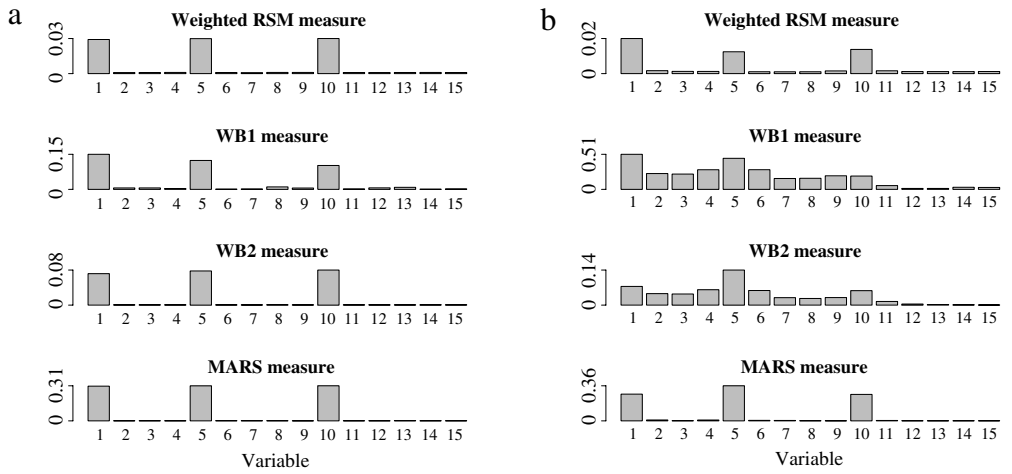

Fig. 5. Normalized weights for model M1 with $M = 1000$ and $n = 200$ based on $L = 200$ simulation trials. Left-hand side figures (a) correspond to $\rho = 0$ whereas right-hand side (b) to $\rho = 0.8$. The weights for first 15 variables are shown in the figure.

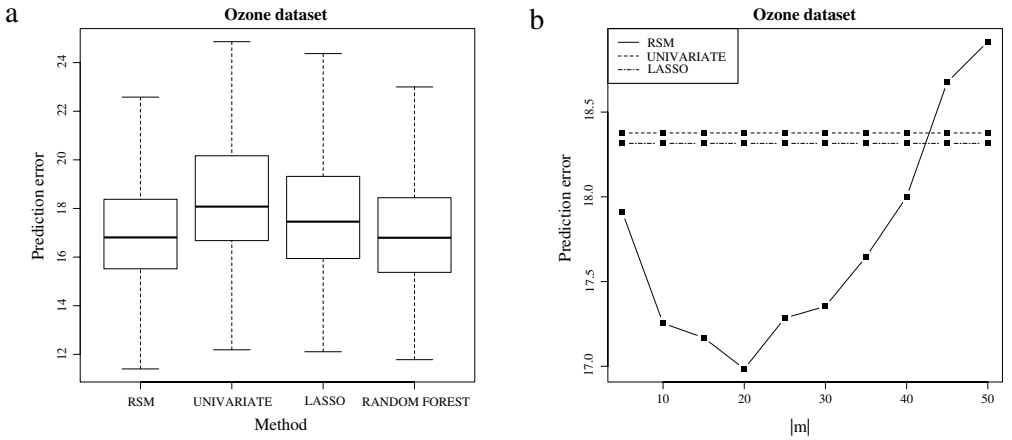

Fig. 6. (a) Prediction errors for Ozone dataset based on 200 random splits of data. (b) Means of prediction errors with respect to $|m|$ for Ozone dataset.

$|m| = \lfloor \min(n_t, M)/2 \rfloor = 27$; for the random forest 1000 regression trees were constructed and the default size of random subspace was equal to $\lfloor M/3 \rfloor = 18$. Results presented in Fig. 6(a) indicate that the RSM performs comparably to the random forest and better than the lasso and the univariate method. Models based on the random forests are more complex when compared with other competitors; an averaged number of variables included in the optimal model was 15.75 (for the RSM), 12.39 (for the univariate method), 16.82 (for the lasso) whereas an averaged number of leaves in the random forests was 59.55. An averaged adjusted coefficient of determination was equal to 0.78 for RSM, and 0.74 for the lasso and the univariate method. Fig. 6(b) presents the means of prediction errors for the RSM against $|m|$, together with analogous quantities for the lasso and the univariate method.

Real data example II

The algorithms are compared on real high-dimensional dataset Mtp2 used in Bergström et al. (2003) who tried to determine whether easily and rapidly calculated 2D and 3D molecular descriptors could predict the melting point of drug-like compounds. The dataset is available at http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html. There are 274 observations representing model drugs and 1143 variables (descriptors). The variables which have less than 5 different values were discarded which resulted in the final number of predictors equal 1041. As before, we split the dataset into three subsets with the sizes of training, validation and test sets equal 100, 100 and 74, respectively. For all considered methods models with intercept were fitted. For the RSM we took $B = 2000$ and $|m| = \min(n_t, M)/2 = 50$; for the random forest 2000 regression trees were constructed based on default number of $M/3 = 347$ randomly chosen predictors. For the lasso method the choice of the optimal parameter α was made using cross-validation on the combined training and validation sets instead of minimizing the prediction error on validation set since the results for the latter method were highly variable. In order to investigate how the methods perform when the number of variables is much greater than 1041 we carried out an

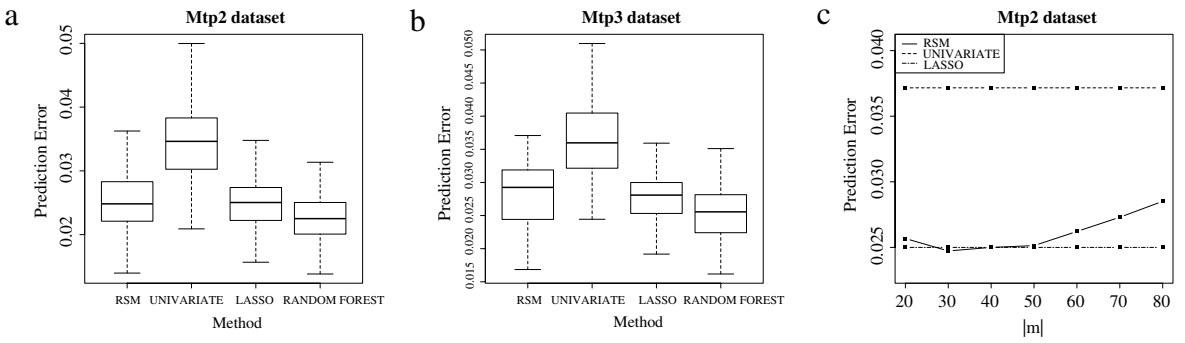

Fig. 7. (a) Prediction errors for Mtp2 dataset based on 200 random splits of data. (b) Prediction errors for Mtp3 dataset based on 200 random splits of data. (c) Means of prediction errors with respect to $|m|$ for Mtp2 dataset.

Table 2
Means of prediction errors (PE) with their standard deviations (Sd) and average number of selected variables (NoV) (leaves in the case of random forests (RF)) for benchmark regression datasets.

Dataset	Method	PE	Sd	NoV
Communities and Crimes ($n = 1994, M = 99$)	RSM	0.021	0.0001	25
	LASSO	0.083	0.001	24
	RF	0.020	0.0008	415
Prostate ($n = 97, M = 8$)	RSM	0.65	0.01	4.5
	LASSO	0.87	0.04	5.3
	RF	0.66	0.01	19
Concrete ($n = 1030, M = 8$)	RSM	0.401	0.002	7.6
	LASSO	0.402	0.001	7.6
	RF	0.122	0.001	203
Forest Fires ($n = 517, M = 12$)	RSM	1.051	0.008	2.2
	LASSO	1.009	0.007	2.2
	RF	1.091	0.008	102
Housing ($n = 506, M = 12$)	RSM	0.3063	0.004	9.5
	LASSO	0.3011	0.004	11.5
	RF	0.1367	0.003	111
Parkinsons ($n = 5875, M = 20$)	RSM	0.757	0.001	15.7
	LASSO	0.753	0.001	16
	RF	0.034	0.0001	1240
Wines ($n = 1599, M = 11$)	RSM	0.667	0.003	8.5
	LASSO	0.661	0.003	9.1
	RF	0.536	0.002	264

additional experiment. Namely $2 \times 1041 = 2048$ additional superfluous explanatory variables were created in two 1041-tuples for each observation by drawing from 1041-dimensional normal distribution with independent components, which mean and variance vector matched that of the original predictors. The total number of predictors thus equals 3123. The new dataset is referred to as Mtp3. Results presented in Fig. 7(a) and (b) indicate that the RSM performs comparably to the lasso, slightly worse than the random forest and better than the univariate method. In both datasets there is no significant difference between the RSM and the lasso when the comparison is based on t-test. However it turns out, that models based on the random forests and the lasso are usually more complex than for RSM; the averaged number of variables included in the optimal model was 11.07 (13.4) for the RSM, 31.71 (42.1) for the lasso whereas an averaged number of leaves in the random forests was 63.04 (62.2) (numbers pertain to Mtp2 and Mtp3 datasets, respectively). An averaged adjusted coefficient of determination was equal to 0.53 (0.48) for the RSM, 0.51 (0.40) for the lasso and 0.16 (0.16) for the univariate method.

Benchmark datasets

The RSM and the lasso were also compared on seven benchmark regression datasets from UCI repository (Frank and Asuncion, 2010). The results presented in Table 2 indicate that for the first two datasets *Communities and Crime* and *Prostate* the RSM outperforms the lasso, for *Forest Fires* and *Parkinsons* datasets the lasso performs better and for the remaining examples both methods perform similarly. Except *Forest Fires* dataset the random forest performs on par or is superior to the RSM and the lasso but at the price of much increased of complexity of the model. The averaged CPU times (secs) on an Intel Core Duo 2.66 GHz processor were for RSM: 8.01 (*Communities and Crimes*), 1.22 (*Prostate*), 1.6 (*Concrete*), 1.41 (*Forest*

Table 3
Positive selection rates, false discovery rates, prediction errors (PE), probabilities of correct ordering (CO) and probabilities of correct selection (CS) based on $L = 200$ simulation trials for model (M1) with $n = 200$, $M = 1000$ and $\rho = 0$.

		$B = 50$	$B = 100$	$B = 250$	$B = 500$
RSM	PSR	0.93	0.99	1	1
	FDR	0.34	0.21	0.11	0.14
	PE	1.33	1.07	1.04	1.04
	CO	0.44	0.77	0.88	0.92
	CS	0.23	0.49	0.68	0.69
WRSM	PSR	1	1	1	1
	FDR	0.11	0.14	0.08	0.08
	PE	1.04	1.04	1.03	1.01
	CO	0.99	1	0.99	1
	CS	0.69	0.66	0.77	0.79

Fires), 1.4 (*Housing*), 5.05 (*Parkinsons*), 1.93 (*Wines*). For datasets having large number of observations (such as *Communities and Crime*, *Parkinsons* and *Wines*) computation times for the random forest were longer than for the RSM. For the lasso the averaged computing time did not exceed 1 s for all datasets.

4.4. Computational considerations

Finally we will discuss the computational cost of the proposed method. Ordering of variables requires $Bn_t|m|^2$ operations when using QR decomposition to fit B linear models with $|m|$ variables each. It follows from the properties of QR decomposition that to fit linear models from the hierarchical list $\{\{i_1\}, \ldots, \{i_1, \ldots, i_{\min(n_t,M)}\}\}$ it suffices to fit only the last one containing $\min(n_t, M)$ variables. This requires $n_t[\min(n_t, M)]^2$ operations which for $|m| = \min(n_t, M)/2$ is smaller than the number of operations for the first step. Applying the RSM to a synthetic dataset generated from M1 for the settings described above with $M = 1000$, $n = 200$, $|m| = 50$ and $B = 1000$ takes an average 4.1 s of CPU time. One drawback of the RSM is that it is highly computer intensive method, which can be overcome by parallel implementation as mentioned in Section 3.

As time complexity of the ordering step is linear in B it is worthwhile to consider variants of the method which would yield similar performance for smaller number of runs. One of the possibilities is a Weighted RSM (WRSM) in which variables are chosen with probabilities proportional to the values of squared t-statistics when univariate models are fitted. Preliminary results, shown in Table 3 for the model M1 indicate that WRSM is superior to ordinary RSM and in this case its performance with $B = 50$ is comparable to the performance of RSM with $B = 500$. Although the gain may be smaller for other models we believe that such variants of the method are worth pursuing.

5. Conclusions

We proposed the random subset method with a new weighting scheme which leads to a novel linear model selection method. It is investigated theoretically and by means of numerical experiments for linear models with a large number of features. We also examined its performance on several real datasets. We have shown in Theorems 2 and 3 that a weight attributed to a variable is asymptotically equivalent to a relative increment of the mean squared error of prediction when the variable is omitted from the model m of a fixed size, averaged over all models m containing it. Numerical experiments for synthetic data sets generated from linear models indicate that the RSM usually works at least on par with the lasso and is frequently superior to it, especially when dependence between predictors is strong or the number of true model variables is small relatively to the number of potential regressors. Moreover, interestingly, the RSM has smaller false discovery rate than the lasso. Similar observations hold for real datasets. Here we also compared both methods to the random forests. For some real datasets, as Mtp2, its extension Mtp3 and some benchmarks considered the random forests yield smaller prediction errors than the RSM but at the price of including many more variables in the model. Although the RSM is much more computationally intensive than the lasso it is not prohibitively so. Its weighted variant WRSM described in Section 4 is less computationally intensive without losing positive features of the original method. Also, numerical experiments for synthetic datasets show that weights of variables pertaining to the RSM yield clear indication of variables contributing to the model from which data is generated.

Acknowledgments

We appreciate insightful comments of two Referees and Associate Editor, which helped to improve significantly the content of the contribution. Research of the second author was partially supported by POKL Project 'Information technologies: research and their interdisciplinary applications'.

Appendix. Proofs

A.1. Proof of Theorem 1

In view of (2) in order to prove that $T^2_{i,m} > T^2_{j,m}$ a.s. it suffices to show that $\mathrm{RSS}(m \setminus \{i\})/n > \mathrm{RSS}(m \setminus \{j\})/n$ a.s. Consider an arbitrary model $s \subset \{1, 2, \ldots, M\}$ such that $\mathbf{X}_s'\mathbf{X}_s$ is invertible. We have the following decomposition

$$n^{-1}\mathrm{RSS}(s) = n^{-1}\boldsymbol{\varepsilon}'(\mathbf{I} - P_s)\boldsymbol{\varepsilon} + n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - P_s)\boldsymbol{\varepsilon} + n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2. \tag{9}$$

First we will show that the first summand converges to σ^2 a.s. In view of the Borel–Cantelli lemma it will follow from $P[|n^{-1}\boldsymbol{\varepsilon}'(\mathbf{I} - P_s)\boldsymbol{\varepsilon} - n^{-1}(n - |s|)\sigma^2| > \epsilon] \le a_n$, for some a_n such that $\sum_n a_n < \infty$. Let χ^2_k be the r.v. having chi-square distribution with k degrees of freedom. We have $P(\chi^2_k \le k - \delta) \le \exp(-\delta^2/4k)$ and $P(\chi^2_k \ge k + \delta) \le \exp(-\delta/4)$, for $\delta > 0$ (see Shibata, 1981). Since $\boldsymbol{\varepsilon}'(\mathbf{I} - P_s)\boldsymbol{\varepsilon}$ has the same distribution as $\sigma^2\chi^2_{n-|s|}$, we have

$$\begin{aligned} P[|n^{-1}\boldsymbol{\varepsilon}'(\mathbf{I} - P_s)\boldsymbol{\varepsilon} - n^{-1}(n - |s|)\sigma^2| > \epsilon] &\le P[\chi^2_{n-|s|} > n - |s| + n\epsilon\sigma^{-2}] + P[\chi^2_{n-|s|} < n - |s| - n\epsilon\sigma^{-2}] \\ &\le \exp(-n\epsilon\sigma^{-2}/4) + \exp(-n^2\epsilon^2\sigma^{-4}/4(n - |s|)), \end{aligned}$$

which implies the desired convergence. Consider now the last summand in (9). In the case of deterministic $\mathbf{X}$ we assumed that limit of $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2$ exists. In the case of random $\mathbf{X}$ we will show that

$$n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2 \xrightarrow{\text{a.s.}} \boldsymbol{\beta}'\Sigma_{\mathbf{x}}\boldsymbol{\beta} - \rho^2_{y,\mathbf{x}_s}\sigma^2_y \tag{10}$$

provided $\mathbf{E}\mathbf{x}_s\mathbf{x}_s'$ is invertible. Let $\mathbf{D}_s$ be an $M \times |s|$ matrix of zeros and ones such that $\mathbf{X}\mathbf{D}_s$ consists of only these $|s|$ columns of $\mathbf{X}$ which correspond to model s, i.e. $\mathbf{X}\mathbf{D}_s = \mathbf{X}_s$. The Law of Large Numbers implies that $n^{-1}\mathbf{X}_s'\mathbf{X}_s \xrightarrow{\text{a.s.}} \mathbf{E}\mathbf{x}_s\mathbf{x}_s'$ as $n \to \infty$ and thus $\mathbf{X}_s'\mathbf{X}_s$ is a.s. invertible for sufficiently large n. Whence the following convergence holds

$$\begin{aligned} n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2 &= n^{-1}\|\mathbf{X}_t\boldsymbol{\beta}_t - P_s\mathbf{X}_t\boldsymbol{\beta}_t\|^2 = n^{-1}(\mathbf{X}_t\boldsymbol{\beta}_t)'[\mathbf{I} - \mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'](\mathbf{X}_t\boldsymbol{\beta}_t) \\ &= n^{-1}(\mathbf{X}_t\boldsymbol{\beta}_t)'(\mathbf{X}_t\boldsymbol{\beta}_t) - n^{-1}\boldsymbol{\beta}_t'\mathbf{X}_t'(\mathbf{X}\mathbf{D}_s)[(\mathbf{X}\mathbf{D}_s)'(\mathbf{X}\mathbf{D}_s)]^{-1}(\mathbf{X}\mathbf{D}_s)'\mathbf{X}_t\boldsymbol{\beta}_t \\ &\xrightarrow{P} \boldsymbol{\beta}_t'\Sigma_{\mathbf{x}_t}\boldsymbol{\beta}_t - \mathrm{cov}(y, \mathbf{x}_s)\Sigma_s^{-1}\mathrm{cov}(\mathbf{x}_s, y) = \boldsymbol{\beta}'\Sigma_{\mathbf{x}}\boldsymbol{\beta} - \boldsymbol{\beta}'\Sigma_{\mathbf{x}}\mathbf{D}_s(\mathbf{D}_s'\Sigma_{\mathbf{x}}\mathbf{D}_s)^{-1}\mathbf{D}_s'\Sigma_{\mathbf{x}}\boldsymbol{\beta}, \end{aligned} \tag{11}$$

where the last convergence follows from $\mathrm{cov}(\mathbf{x}_s, \mathbf{x}_t)\beta_t = \mathrm{cov}(\mathbf{x}_s, y)$ as $y = \mathbf{x}'\beta + \varepsilon$. Convergence in (10) now follows from (5). Consider the second term in (9). Provided that $\mathbf{X}_s'\mathbf{X}_s$ is invertible, $n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - P_s)\boldsymbol{\varepsilon}$ given $\mathbf{X}$ has $N(0, v_n)$ distribution, where $v_n = n^{-2}4\sigma^2\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2$. Since the limit of a sequence $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2$ exists (in the case of random $\mathbf{X}$ with probability 1) then $v_n = O(n^{-1})$ a.s. Let Z be a random variable with standard normal distribution and ϕ be its density function. Using Mill's inequality we have

$$P[|n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - P_s)\boldsymbol{\varepsilon}| > \delta|\mathbf{X}] = P(|Z| > \delta v_n^{-1/2}) \le 2\delta^{-1}v_n^{1/2}\phi(\delta v_n^{-1/2}),$$

for $\delta > 0$, which yields the convergence $n^{-1}2(\mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - P_s)\boldsymbol{\varepsilon} \xrightarrow{\text{a.s.}} 0$ in the case of deterministic $\mathbf{X}$. In the case of random $\mathbf{X}$ the proof is analogous. The assertions of the theorem follow from (9) and the above reasoning taking as s models $m \setminus \{i\}$ and $m \setminus \{j\}$ respectively. Needed invertibility of $\Sigma_{m\setminus\{i\}}$ follows form that of Σ_m and Sylvester's theorem. Note that it follows from the proof that in the case of random $\mathbf{X}$ almost surely

$$\frac{\mathrm{RSS}(s)}{n} \to \sigma^2 + \boldsymbol{\beta}'\Sigma_{\mathbf{x}}\boldsymbol{\beta} - \rho^2_{y,\mathbf{x}_s}\sigma^2_y = \sigma^2_y - \mathrm{var}(P_s y) = \mathrm{var}(y - P_s y), \tag{12}$$

where the penultimate equality follows from (4) and the last one from the orthogonality of $P_s y$ and $y - P_s y$.

A.2. Proof of Corollary 1

In the case $m \supseteq t$ we have $n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_{m\setminus\{i\}}\mathbf{X}\boldsymbol{\beta}\|^2 = 0$, for $i \notin t$. Thus the assertion (i) follows easily from the proof of (9) and assumption (8). In the case of random $\mathbf{X}$ we will show that (8) holds a.s., i.e.

$$n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_{m\setminus\{i\}}\mathbf{X}\boldsymbol{\beta}\|^2 \xrightarrow{\text{a.s.}} \lambda_{m\setminus\{i\}} > 0, \tag{13}$$

for $i \in t$. Consider model $s = m \setminus \{i\} \not\supseteq t$. Matrix $\Sigma_{\mathbf{x}_m}$ as a positive definite matrix can be decomposed as $\Sigma_{\mathbf{x}_m} = \Sigma^{1/2}_{\mathbf{x}_m}\Sigma^{1/2}_{\mathbf{x}_m}$ where $\Sigma^{1/2}_{\mathbf{x}_m} = \mathbf{U}\Xi^{1/2}\mathbf{U}'$, $\mathbf{U}$ is an orthogonal matrix and Ξ is a diagonal matrix with positive diagonal. Then using (11) and denoting by $\boldsymbol{\beta}_m$ the restriction of $\boldsymbol{\beta}$ to m we have in view of $m \supseteq t$

$$\begin{aligned} \lambda_s &= \boldsymbol{\beta}_m'\Sigma_{\mathbf{x}_m}\boldsymbol{\beta}_m - \boldsymbol{\beta}_m'\Sigma_{\mathbf{x}_m}\mathbf{D}_s(\mathbf{D}_s'\Sigma_{\mathbf{x}_m}\mathbf{D}_s)^{-1}\mathbf{D}_s'\Sigma_{\mathbf{x}_m}\boldsymbol{\beta}_m \\ &= (\Sigma^{1/2}_{\mathbf{x}_m}\boldsymbol{\beta}_m)'[\mathbf{I} - \Sigma^{1/2}_{\mathbf{x}_m}\mathbf{D}_s(\mathbf{D}_s'\Sigma_{\mathbf{x}_m}\mathbf{D}_s)^{-1}\mathbf{D}_s'\Sigma^{1/2}_{\mathbf{x}_m}](\Sigma^{1/2}_{\mathbf{x}_m}\boldsymbol{\beta}_m) = \|(\Sigma^{1/2}_{\mathbf{x}_m}\boldsymbol{\beta}_m) - Q_s(\Sigma^{1/2}_{\mathbf{x}_m}\boldsymbol{\beta}_m)\|^2 > 0, \end{aligned}$$

where Q_s is a projection on the column space spanned by the columns of $\Sigma^{1/2}_{\mathbf{x}_m}$ corresponding to model s. The last inequality follows from the fact that the columns of $\Sigma^{1/2}_{\mathbf{x}_m}$ are linearly independent and model s does not contain at least one significant variable. Thus (13) is obtained.

A.3. Proof of Corollary 2

Let $\Sigma_{\mathbf{x}} = (\sigma_{ij})$. Consider random $\mathbf{X}$ case and some model s such that $\Sigma_{\mathbf{x}_{s\cup t}}$ is diagonal and invertible. Then it follows from (11)

$$n^{-1}\|\mathbf{X}\boldsymbol{\beta} - P_s\mathbf{X}\boldsymbol{\beta}\|^2 \xrightarrow{\text{a.s.}} \boldsymbol{\beta}'\Sigma_{\mathbf{x}}\boldsymbol{\beta} - \boldsymbol{\beta}'\Sigma_{\mathbf{x}}\mathbf{D}_s(\mathbf{D}_s'\Sigma_{\mathbf{x}}\mathbf{D}_s)^{-1}\mathbf{D}_s'\Sigma_{\mathbf{x}}\boldsymbol{\beta} = \boldsymbol{\beta}'\Sigma_{\mathbf{x}}\boldsymbol{\beta} - \sum_{k\in s\cap t}\sigma_{kk}\beta_k^2. \tag{14}$$

Since $\sigma_{kk} > 0$, for $k \in m$ it follows that for $i \in t \cap m$ and $j \in m \cap t^c$

$$\sum_{k\in\{m\setminus\{j\}\}\cap t}\sigma_{kk}\beta_k^2 > \sum_{k\in\{m\setminus\{i\}\}\cap t}\sigma_{kk}\beta_k^2,$$

which together with (14) implies the assertion. The proof is analogous in the case of deterministic $\mathbf{X}$.

A.4. Proof of Theorem 2

First note that

$$\mathbf{E}^*\frac{T_{i,m^*}^2}{n-|m|} = \frac{1}{|\mathcal{M}_{|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}}\frac{T_{i,m}^2}{n-|m|} \tag{15}$$

and for almost any sequence $(\mathbf{Y}_n, \mathbf{X}_n)_{n=1}^{\infty}$

$$\begin{aligned}\mathrm{Var}^*\frac{T_{i,m^*}^2}{n-|m|} &= \frac{1}{|\mathcal{M}_{|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}}\frac{T_{i,m}^4}{(n-|m|)^2} - \left(\frac{1}{|\mathcal{M}_{|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}}\frac{T_{i,m}^2}{n-|m|}\right)^2 \\ &\to \frac{1}{|\mathcal{M}_{|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}} t_{i,m}^2 - \left(\frac{1}{|\mathcal{M}_{|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}} t_{i,m}\right)^2 < \infty, \quad \text{as } n\to\infty.\end{aligned} \tag{16}$$

Using (15), (16) and Markov's inequality we have that

$$\frac{1}{B}\sum_{m^*:i\in m^*}\frac{T_{i,m^*}^2}{n-|m|} - \mathbf{E}^*\frac{T_{i,m^*}^2}{n-|m|} \xrightarrow{P^*} 0, \quad \text{as } n\to\infty.$$

Thus using the fact that $\frac{C_{i,B_n}}{B_n} \xrightarrow{P^*} \frac{|\mathcal{M}_{i,|m|}|}{\mathcal{M}_{|m|}}$ we obtain

$$\mathrm{TS}_i^* - \frac{1}{|\mathcal{M}_{i,|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}}\frac{T_{i,m}^2}{n-|m|} \xrightarrow{P^*} 0, \quad \text{as } n\to\infty,$$

which, together with $(n-|m|)^{-1}T_{i,m}^2 \to t_{i,m}$ for almost any sequence $(\mathbf{Y}_n, \mathbf{X}_n)_{n=1}^{\infty}$, yields the assertion of the theorem.

A.5. Proof of Corollary 3

Consider some model m and $i \in m$. It follows from the proof of Corollary 2 that

$$t_{i,m} = \frac{\sigma_{ii}\beta_i^2}{\sigma^2 + \sum_{k\in(m\cap t)^c}\sigma_{kk}\beta_k^2}.$$

Thus from Theorem 2 we have

$$\mathrm{TS}_i^* \xrightarrow{P^*} \frac{1}{|\mathcal{M}_{i,|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}}\frac{\sigma_{ii}\beta_i^2}{\sigma^2 + \sum_{k\in(m\cap t)^c}\sigma_{kk}\beta_k^2} > 0,$$

for $i \in t$ and $\mathrm{TS}_i^* \xrightarrow{P^*} 0$ for $i \notin t$ which yields the assertion.

A.6. Proof of Theorem 3

Consider first the case of deterministic **X**. It follows form the proof of Theorem 2 that as (iii) is satisfied, $\mathrm{Var}(T_{i,m*}/(n-|m|)) = O_{P*}(1)$ and thus it is enough to check that $\max |(n-|m|)^{-1}T_{i,m}^2 - t_{i,m}| \to 0$ a.s., where the maximum is taken over subsets m of the fixed size $|m|$. This in its turn in view of (2) follows from

$$\max_{s\subset\{1,\ldots,M_n\}} |n^{-1}\mathrm{RSS}(s) - (\sigma^2+\lambda_s)| \to 0 \quad \text{a.s.}, \tag{17}$$

with s being a subset of the fixed size $|m|$ or $|m|-1$. We outline the proof of (17) for $|s|=|m|$, the other case is analogous. Consider decomposition (9) of $n^{-1}\mathrm{RSS}(s)$. Reasoning as in the proof of Theorem 1 we easily obtain that for $|s|=|m|$

$$P(\max_s |n^{-1}\boldsymbol{\varepsilon}'(\mathbf{I}-P_s)\boldsymbol{\varepsilon} - \sigma^2| \geq \epsilon) = O\left(\binom{M_n}{|m|}\exp(-n\epsilon)\right),$$

which is summable in view of (ii). Analogous reasoning establishes uniform (in s) convergence to 0 of the second term in (9). The third term converges uniformly in view of assumption (i). In the case of random **X** we prove that (i′) and (ii′) imply that (i) holds a.s. This relies on

$$\max_s |n^{-1}\mathbf{X}_s'\mathbf{X}_s - \Sigma_{\mathbf{x}_s}| \to 0 \quad \text{a.s.} \tag{18}$$

Similar result has been proved by Bühlmann (2006) for uniformly bounded x_j, $j=1,\ldots,M_n$. The proof in the more general case (ii′) uses Bernstein's inequality and the second part of (ii′) for truncated r.v.s. $x_j^{T_n} = x_j\mathbf{I}(|x_j|\leq T_n) + T_n\mathrm{sgn}(x_j)\mathbf{I}(|x_j|>T_n)$ where $T_n = n^{1/4-\gamma}$ with sufficiently small $\gamma > 0$. We omit the details. Moreover it follows from (i′) that $\det(\Sigma_{\mathbf{x}_s})$ is uniformly bounded away from 0. Thus in view of (18) we have that

$$\max_s \left|\left(\frac{\mathbf{X}_s'\mathbf{X}_s}{n}\right)^{-1} - \Sigma_{\mathbf{x}_s}^{-1}\right| \to 0 \quad \text{a.s.} \tag{19}$$

and consequently in the view of (18) and (19) the third term in (9) tends to 0 uniformly a.s. The reasoning for the two remaining terms (9) is analogous to the proof of deterministic case.

A.7. Proof of Corollary 4

From Theorem 3 we have that $\mathrm{TS}_i^* - ts_i \xrightarrow{P^*} 0$. Taking into account the proof of Corollary 3 we have that

$$\liminf_n ts_i = \liminf_n \frac{1}{|\mathcal{M}_{i,|m|}|}\sum_{m\in\mathcal{M}_{i,|m|}} \frac{\sigma_{ii}\beta_i^2}{\sigma^2 + \sum_{k\in(m\cap t)^c}\sigma_{kk}\beta_k^2} > 0,$$

for $i\in t$. Moreover, for any finite I such that $I\cap t = \emptyset$ $\max_{i\in I}\mathrm{TS}_i^* \xrightarrow{P^*} 0$ from which the conclusion follows.

References

Bergström, C., Norinder, U., Luthman, K., Artursson, P., 2003. Molecular descriptors influencing melting point and their role in classification of solid drugs. Journal of Chemical Information and Computer Sciences 43 (4), 1177–1185.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Friedman, J.H., 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association 80 (391), 580–598.

Bühlmann, P., 2006. Boosting for high-dimensional linear models. Annals of Statistics 34, 559–583.

Bühlmann, P., Kalisch, M., Maathuis, M.H., 2010. Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. Biometrika 97, 261–278.

Casella, G., Giron, J., Martinez, M.L., Moreno, E., 2009. Consistency of Bayesian procedures for variable selection. Annals of Statistics 37, 1207–1228.

Chevan, A., Sutherland, M., 1991. Hierarchical partitioning. The American Statistician 45, 90–96.

Donoho, D.L., 2000. High-dimensional data analysis: the curses and blessings of dimensionality. In: Aide-Memoire of a Lecture at AMS Conference on Math Challenges of the 21st Century.

Draminski, M., Rada-Iglesias, A., Enroth, S., Wadelius, C., Koronacki, J., Komorowski, J., 2008. Monte Carlo feature selection for supervised classification. Bioinformatics 24 (1), 110–117.

Fan, J., Lv, J., 2008. Sure independence screening for ultra-high dimensional feature space (with discussion). Journal of the Royal Statistical Society. Series B 70, 849–911.

Feldman, B., 1999. Relative importance and value. Unpublished Manuscript.

Frank, A., Asuncion, A., 2010. UCI machine learning repository. URL: http://archive.ics.uci.edu/ml.

Frommlet, F., Ruhaltinger, F., Twarog, B., Bogdan, M., 2012. Modified versions of Bayesian information criterion for genome-wide association studies. Computational Statistics and Data Analysis 56, 1038–1051.

Grömping, U., 2007. Estimators of relative importance in linear regression based on variance decomposition. The American Statistician 61, 139–147.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8), 832–844.

Huang, J., Ma, S., Zhang, C.-H., 2008. Adaptive lasso for high-dimensional regression models. Statistica Sinica 18, 1603–1618.
Kuhn, M., 2008. Building predictive models in R using the caret package. Journal of Statistical Software 28 (5), 1–26.
Lai, C., Reinders, M.J.T., Wessels, L., 2006. Random subspace method for multivariate feature selection. Pattern Recognition Letters 27, 1067–1076.
Lindemann, R., Merenda, P., Gold, R., 1980. Introduction to Bivariate and Multivariate Analysis. Scott, Foresman, Glenview.
Sandri, M., Zuccolotto, P., 2010. Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. Statistics and Computing 20, 393–407.
Shao, J., 1993. Linear model selection by cross-validation. Journal of the American Statistical Association 88, 486–494.
Shao, J., Deng, X., 2012. Estimation in high-dimensional linear models with deterministic covariates. Annals of Statistics 40 (2), 812–831.
Shibata, R., 1981. An optimal selection of regression variables. Biometrika 63, 117–126.
Stoppiglia, H., Dreyfus, G., Dubois, R., Oussar, Y., 2003. Ranking a random feature for variable and feature selection. Journal of Machine Learning Research 3, 1399–1414.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B 58, 267–288.
Zhang, P., 1992. On the distributional properties of model selection criteria. Journal of the American Statistical Association 87, 732–737.
Zheng, X., Loh, W.-Y., 1995. Consistent variable selection in linear models. Journal of the American Statistical Association 90 (429), 151–156.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B 67 (2), 301–320.