Selection and Prediction for Linear Models using Random Subspace Methods

Jan Mielniczuk[1,2] and Paweł Teisseyre[1]

[1] Institute of Computer Science, Polish Academy of Sciences,
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

[2] Warsaw University of Technology, Faculty of Mathematics and Information Science,
ul. Koszykowa 75, 00-662 Warsaw, Poland

Abstract

In the paper two versions of random subspace method (RSM) for linear regression models are considered. In the original RSM in regression framework introduced in [1] subsets of regressors are sampled with equal probabilities from all subsets of a chosen size and then fitted in order to construct scores of all variables. Proposed modifications consist in preferential sampling of variables according to preliminary assessment of their importance and/or initial screening of features. Some properties of the proposed methods are discussed and their performance as prediction method for moderate sample sizes is studied by means of simulations. The first variant, weighted RSM, behaves promisingly when the dependence between regressors is not very strong and is also much less computationally expensive than the RSM.

1 Introduction

Model selection in high dimensional feature space plays an important role in diverse fields of sciences, engineering and humanities. Examples include microarray analysis, Quantitative Trait Loci (QTL) analysis, Genome-Wide Association Study (GWAS), drug design analysis and high-frequency financial analysis among others. In such problems it is challenging to find important variables out of thousands of predictors, with number of observations usually in tens or hundreds. In [2] the need for development of high-dimensional data analysis is discussed. Since the true relationship in data is usually unknown, very often it is worthwhile to include higher degree terms as well as interaction terms to the model. This can substantially increase the number of potential attributes. The problem recently has received much attention in the statistical and machine learning literature. An intensively studied line of research is focused on regularization (cf. e.g. [3], [4]). In many approaches a preliminary feature selection is used, e.g. in [5] a method of dimensionality reduction based on so called sure independence screening is proposed. Let us also mention procedures using information criteria modified to high-dimensional setup, see e.g. [6] or [7]. Recently a novel approach based on the adaptation of the random subspace method (RSM) in the regression context has been proposed in [1].

In the RSM a random subset m with $|m|$ features, smaller than the number of all predictors p and a number of observations n, is chosen and the model is fitted in the reduced feature space. Selected features are assigned weights describing their relevance in the considered submodel. In order to cover a large portion of features in the dataset, the selection is repeated B times and the cumulative weights (called final scores) are computed. The results of all iterations are combined in a list of p features ordered according to final scores. The final model can be constructed based on predetermined number of the most significant variables or using selection method applied to the nested list of models given by the ordering. The simulation experiments described in [1] 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. One drawback of this method is its computational cost. When the number of features p is large we should take large B in order to ensure that all variables are likely to be selected to random subspaces.

In this paper we propose two modifications of the original algorithm. In the first method, called weighted random subspace method (WRSM), variables are chosen to subspaces with probabilities proportional to the values of individual weights when univariate models are fitted. In the second method, called screened random subspace method (SRMS), the preliminary feature screening is performed. Both approaches reduce the computational cost of the original procedure.

This paper is organized as follows. The original RSM algorithm is recalled in Section 2.1; the choice of the weights is described in Section 2.2 and in Section 2.3 some additional properties of quantities related to the weights are discussed. The modifications of the RSM are presented in Section 2.4 and the results of numerical experiments are discussed in Section 3. The proofs are relegated to the appendix.

We define now the formal setup of the paper. Let $(\mathbf{Y},\mathbf{X})$ be the observed data, where $\mathbf{Y}=\mathbf{Y}^{(n)}$ is an $n\times 1$ vector of n responses whose variability we would like to explain and $\mathbf{X}=\mathbf{X}^{(n)}$ is a $n\times p$ design matrix consisting of vectors of p 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 $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)'$ is an unknown vector of parameters. We consider two scenarios: the case of deterministic and random $\mathbf{X}$. In the latter case rows of $\mathbf{X}$ constitute n independent realizations of p-dimensional random variable $\mathbf{x}$ and coordinates of vector $\mathbf{Y}$ form an i.i.d. sample distributed as $y=\mathbf{x}'\boldsymbol{\beta}+\varepsilon$. A distribution of $\mathbf{x}=(x_1,\ldots,x_p)'$ 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. The number of attributes p may be larger than n. As any submodel of (1) containing $|m|$ variables $(x_{i_1},\ldots,x_{i_{|m|}})'$ can be described by set of indices $m=\{i_1,\ldots,i_{|m|}\}$ in order to make notation simpler it will be referred to as model m. We denote by $\mathbf{X}_m$ the matrix composed of the columns of $\mathbf{X}$ with indices in m and by $\mathbf{x}_m$ a subvector of $\mathbf{x}$ consisting of coordinates corresponding to m. Similarly, $\boldsymbol{\beta}_m\in R^{|m|}$ denotes the vector consisting of components of $\boldsymbol{\beta}$ with indices in m. For simplicity, model fitted to data $(\mathbf{Y},\mathbf{X}_m)$ we will be denoted by $y\sim\mathbf{x}_m$. Usually 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 $\{i : \beta_i \neq 0\}$ 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 $t \subset \{1, 2, \ldots, p\}$ is unique and t does not change with n.

2 Random Subspace Methods

2.1 Main Algorithm

We first describe the basic algorithm of Random Subspace Method.

RSM Algorithm

1. Input: observed data $(\mathbf{Y}, \mathbf{X})$, number of subset draws B, size of the subspace $|m| < \min(p, n)$. Choice of weights $w_{i,m}$ is described in Section 2.2.
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 weight w_{i,m^*} for each $i \in m^*$. Set $w_{i,m^*} = 0$ 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 score FS_i^* defined as
$$FS_i^* = \frac{1}{C_{i,B}} \sum_{m^* : i \in m^*} w_{i,m^*}.$$
4. Sort the list of variables according to scores FS_i^*: $FS_{i_1}^* \geq FS_{i_2}^* \ldots \geq FS_{i_p}^*$.
5. Output: Ordered list of variables $\{i_1, \ldots, i_p\}$.

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 dependencies 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 limited by $\min(n, p)$. In the following the value of parameter $|m|$ is chosen empirically. We concluded from numerical experiments that the reasonable choice is $|m| = \min(n, p)/2$. It follows from the description above that a parallel version of the algorithm is very easy to implement.

2.2 Choice of the weights $w_{i,m}$

In this section we discuss rationale for using a squared value of t-statistic as a weight in RSM procedure. Observe first that a randomly chosen model m in the second step of RSM procedure may be misspecified, in the sense that it may not contain all significant variables. Thus it is important to investigate the performance of proposed weights in a general case when a considered model may be wrong. 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 discuss the problem in Theorem 1 which states the conditions under which such a situation can not 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 $|m|<\min(n,p)$. 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

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

be t-statistic corresponding to variable i when model m is fitted to the data. In the above formula $\hat{\sigma}^2_m=(n-|m|)^{-1}RSS(m)$, where $RSS(m)=\mathbf{Y}'(\mathbf{I}-P_m)\mathbf{Y}$ is 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 this model. The following equality holds

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

Thus $T^2_{i,m}/(n-|m|)$ is a relative increase of RSS when variable i is dropped from the model m. It follows from (2) and generalized Cochran theorem that $T^2_{i,m}/(n-|m|)$ is a ratio of two independent chi squared distributed random variables: $\chi^2_1(\lambda_1)$ in the case of numerator and $\chi^2_{n-|m|}(\lambda_2)$ for denominator, where parameters of noncentrality are equal $\lambda_1=||(P_m-P_{m\setminus\{i\}})\mathbf{X}\beta||^2/(2\sigma^2)$ and $\lambda_2=||(I-P_m)\mathbf{X}\beta||^2/(2\sigma^2)$, respectively. It will be shown is Section 2.3 that λ_2 is equal to the Kullback–Leibler divergence between probability density function corresponding to the true model and space spanned by columns of $\mathbf{X}$ corresponding to model m. Note also that due to a variance decomposition for a linear model which includes constant regressor we have

$$\frac{T^2_{i,m}}{n-|m|}=\frac{R^2_m-R^2_{m\setminus\{i\}}}{1-R^2_m}, \tag{3}$$

where R^2_m is a coefficient of determination for a model m. Equation (3) provides the main motivation for our choice of weights in RSM scheme, that is we consider $w_{i,m}=(n-|m|)^{-1}T^2_{i,m}$. Namely, it indicates that up to a multiplicative factor, $T^2_{i,m}$ is a decrease in R^2 due to leaving out x_i multiplied by a measure of goodness-of-fit $(1-R^2_m)^{-1}$ 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 $i \in \{1, \ldots, p\}$. 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

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

be the squared multiple correlation coefficient between y and its projection on a subspace spanned by coordinates of $\mathbf{x}_m$. It is easy to see that

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

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 $\rho^2_{y,\mathbf{x}_m}$ equals the maximal value of a squared correlation between y and linear combination of coordinates 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.

Let $\lambda_n(m) := ||\mathbf{X}\boldsymbol{\beta} - P_m\mathbf{X}\boldsymbol{\beta}||^2$. In the case of deterministic $\mathbf{X}$ let

$$\lambda(m) := \lim_{n\to\infty} n^{-1}\lambda_n(m).$$

For random $\mathbf{X}$ the limit is understood almost surely. Note that $\lambda_n(m)$ 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. Since $\lambda_n(m)$ is an important object we discuss its properties in Section 2.3. Proposition 1 below gives an interpretation of $\lambda(m)$ in the 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 $\lambda(m \setminus \{i\})$ exists almost surely and the ordering can be reexpressed in the 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 $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_t\boldsymbol{\beta}_t$, $\lambda(m)$ does not depend on the number of potential regressors p. The same observation applies to $T^2_{i,m}$. The following results have been proved in [1].

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 $T^2_{i,m} \geq T^2_{j,m}$ almost surely for sufficiently large n implies

$$\lambda(m \setminus \{i\}) \geq \lambda(m \setminus \{j\}). \tag{6}$$

Moreover, strict inequality in (6) implies $T^2_{i,m} > T^2_{j,m}$ almost surely for sufficiently large n.
(ii) In the case of random $\mathbf{X}$ assume that $\Sigma_{\mathbf{x}_m}$ is invertible and $\mathbf{E}x^4_j$ are finite for all $j \in m$. Then $T^2_{i,m} \geq T^2_{j,m}$ almost surely for sufficiently large n implies

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

Moreover, strict inequality in (7) implies $T^2_{i,m} > T^2_{j,m}$ almost surely for sufficiently large n.

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). It is also easy to see that for m having two elements condition (7) is equivalent to $\rho^2(y, x_i) = \rho^2(y, \mathbf{x}_{m\setminus\{j\}}) > \rho^2(y, \mathbf{x}_{m\setminus\{i\}}) = \rho^2(y, x_j)$.

Proposition 1 *When $\mathbf{X}$ is deterministic consider the mean squared error of prediction for OLS estimation in model m*

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

where $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$ with $\boldsymbol{\varepsilon}^*$ being an independent copy of $\boldsymbol{\varepsilon}$. Let

$$MSEP(m) = lim_{n\to\infty} n^{-1} MSEP_n(m).$$

Thus the ordering in (6) is equivalent to ordering

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

Moreover, for random $\mathbf{X}$, (7) is equivalent to

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

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 condition

$$\lambda(m \setminus \{i\}) > 0, \tag{8}$$

for all $i \in t$ implies that $\min_{i\in t} T^2_{i,m} > \max_{i\in t^c\cap m} T^2_{i,m}$ almost surely for sufficiently large n.
(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.

Corollary 1 asserts that when $m \supseteq t$ the relevant variables precede the spurious ones asymptotically provided that column $\mathbf{X}_i$ for any $i \in t$ is separated from the linear space spanned by other columns in m. Below we provide simple sufficient condition for (8) to hold.

Proposition 2 *For deterministic X and $m \not\supseteq t$ assume that $n^{-1}\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m} \to W$, as $n \to \infty$, W is positive definite matrix. Then $\lambda(m) > 0$.*

The proof of Proposition 2 is relegated to Appendix. Various versions of condition (8) are used to prove asymptotic results of model selection for linear models (cf [8], [9], [10], [11]). 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 Bayes selection method introduced there. Note the fact that (8) is automatically satisfied for random $\mathbf{X}$ which can be regarded as 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 a 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. Example 1 explores such a situation.

Example 1 Consider random-design regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta} = (\beta_1, 0, \beta_3)'$, $\boldsymbol{\varepsilon}$ has $N(0, \mathbf{I})$ distribution and rows of $\mathbf{X}$ are normally distributed with covariance matrix

$$\Sigma_{\mathbf{x}} = (\sigma_{ij}) = \begin{bmatrix} 1 & b & 0 \\ b & 1 & a \\ 0 & a & 1 \end{bmatrix},$$

where $a, b \in (0,1)$ are parameters. Thus the true variables are uncorrelated, correlation between true variable x_1 and spurious x_2 is equal b whereas correlation between true variable x_3 and spurious x_2 is equal a. A misspecified model $m = \{1,2\}$ containing two variables only is fitted: x_1 (true) and x_2 (spurious). Theorem 1 (ii) states that $T^2_{1,m} > T^2_{2,m}$ for sufficiently large n with probability 1 i.e. the true 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

$$\sigma_{11}^{-1}(\beta_1\sigma_{11} + \beta_2\sigma_{12} + \beta_3\sigma_{13})^2 > \sigma_{22}^{-1}(\beta_1\sigma_{12} + \beta_2\sigma_{22} + \beta_3\sigma_{23})^2$$

or equivalently $\rho^2(x_1, y) > \rho^2(x_2, y)$. For $\beta_1 = \beta_3 = 1$ an easy calculation shows this is equivalent to $1 > b + a$. When the spurious variable x_2 is strongly correlated with true ones it takes over their roles in the misspecified model and in effect has more predictive power than variable x_1. For $\beta_1 = \beta_3 = 1$ we carried out $L = 500$ simulations for $n = 100, 200, 500$ and computed fraction of correct orderings for which $T^2_{1,m} > T^2_{2,m}$ with varying value of parameter a and for fixed $b = 0.5$. The results are presented in Figure 1. Note that to the left of the value $a = 0.5$ probability of correct ordering significantly increases in concordance with the condition $a + b < 1$. When the correlation a between spurious variable x_2 and true variable x_3 missing from the model is strong then the ordering of variables in m induced by t-statistics can be incorrect with high probability, i.e. it is likely that $T^2_{1,m} < T^2_{2,m}$. Note that when model $m = \{2,3\}$ is fitted the condition for correct ordering is the same.

2.3 Properties of $\lambda_n(m)$

In this section we discuss some formal properties of term $\lambda_n(m)$ defined above in Theorem 1. Proposition 3 below gives an interpretation of $\lambda_n(m)$ in terms of the Kullback–Leibler divergence (KL). Let $\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}(s)$ be the probability density function (p.d.f.) of conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$, i.e the p.d.f. of the multivariate

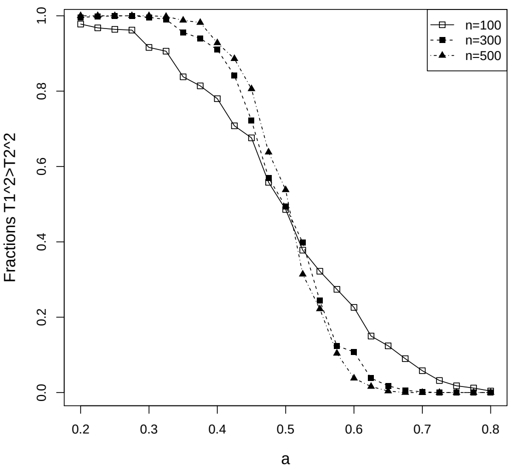


FIGURE 1: Estimated probabilities of $T^2_{1,m} > T^2_{2,m}$ with respect to a based on $N = 500$ trials.

normal distribution $N(\mathbf{X}_t\boldsymbol{\beta}_t, \sigma^2\mathbf{I})$. Let $\mathbf{f}_{\mathbf{X}_m\boldsymbol{\beta}_m}(s)$ be the p.d.f. corresponding to model m, i.e. the p.d.f. of $N(\mathbf{X}_m\boldsymbol{\beta}_m, \sigma^2\mathbf{I})$. Let $sp_{\mathbf{X}}(m)$ denote the space spanned by columns of $\mathbf{X}$ corresponding to model m. The following properties hold. They are proved in the appendix.

Proposition 3

$$KL(\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}, \mathbf{f}_{\mathbf{X}_m\boldsymbol{\beta}_m}) = \int_{-\infty}^{+\infty} \mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}(s) \log \frac{\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}(s)}{\mathbf{f}_{\mathbf{X}_m\boldsymbol{\beta}_m}(s)} ds = \frac{||\mathbf{X}_t\boldsymbol{\beta}_t - \mathbf{X}_m\boldsymbol{\beta}_m||^2}{2\sigma^2} \qquad (9)$$

and

$$KL(\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}, sp_{\mathbf{X}}(m)) = \inf_{\mathbf{X}_m\boldsymbol{\gamma}\in sp_{\mathbf{X}}(m)} KL(\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}, \mathbf{f}_{\mathbf{X}_m\boldsymbol{\gamma}}) = \frac{\lambda_n(m)}{2\sigma^2}. \qquad (10)$$

It follows that $\lambda(m)$ is equal, up to multiplicative factor $(2\sigma^2)^{-1}$, to a limiting value of Kullback–Leibler divergence, averaged per observation, between probability density function corresponding to the true model and space spanned by columns of $\mathbf{X}$ corresponding to model m.

Proposition 4 *The following equality holds*

$$\lambda_n(m) = \boldsymbol{\beta}'_{t\backslash m}[\mathbf{X}'_{t\backslash m}\mathbf{X}_{t\backslash m} - \mathbf{X}'_{t\backslash m}\mathbf{X}_m(\mathbf{X}'_m\mathbf{X}_m)^{-1}\mathbf{X}'_m\mathbf{X}_{t\backslash m}]\boldsymbol{\beta}_{t\backslash m}.$$

It is seen that the matrix pertaining to quadratic form above is Schur complement (see e.g. [12], p. 95) of the block $\mathbf{X}'_m\mathbf{X}_m$ of the matrix $\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m}$.

Proposition 5 *The following inequality holds*

$$\lambda_n(m) \geq \lambda_{\min}(\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m})||\boldsymbol{\beta}_{t\backslash m}||^2.$$

It follows from Proposition 5 that $\lambda_{\min}(\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m}) > 0$ implies that $\lambda_n(m)$ is also positive.

2.4 Modifications of the original approach

As time complexity of the calculation of final scores is linear in B it is worthwhile to consider variants of the method which would yield similar performance for smaller number of runs. Here we introduce two algorithms: Weighted RSM (WRSM) and Screened RSM (SRSM), which also can be combined together. First we will describe the WRSM procedure.

WRSM Algorithm

1. Input: observed data $(\mathbf{Y}, \mathbf{X})$, number of subset draws B, size of the subspace $|m| < \min(p, n)$.
2. For each variable i fit univariate model $y \sim x_i$ and compute weight of i-th variable $w_i^{(0)}$.
3. For each variable i compute $\pi_i = w_i^{(0)} / \sum_{l=1}^{p} w_l^{(0)}$.
4. 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 in such a way that probability of choosing i-th variable is equal π_i.
    - Fit model $y \sim \mathbf{x}_{m^*}$ and compute weight w_{i,m^*} for each $i \in m^*$. Set $w_{i,m^*} = 0$ if $i \notin m^*$.
    - Update the counter $C_{i,k} = C_{i,k-1} + I\{i \in m^*\}$.
5. For each variable i compute the final score FS_i^* defined as
$$FS_i^* = \frac{1}{C_{i,B}} \sum_{m^* : i \in m^*} w_{i,m^*}.$$
6. Sort the list of variables according to scores FS_i^*: $FS_{i_1}^* \geq FS_{i_2}^* \ldots \geq FS_{i_p}^*$.
7. Output: Ordered list of variables $\{i_1, \ldots, i_p\}$.

Actual point 4 of the above procedure uses a simplified scheme, namely probabilities π_i are applied sequentially, that is the probability of choosing the next variable is proportional to the probabilities amongst variables not chosen till that moment. Note that this does not match exactly the procedure given in the algorithm. Probability that the given i-th variable will be selected to a randomly drawn model m^* is

$$\begin{aligned}&P(i \in m^*) = P(i \text{ be selected in the first step}) + \\ &P(i \text{ be selected in the second step}) + \ldots + P(i \text{ be selected in the } m\text{-th step}) = \\ &\pi_i + \pi_i \sum_{j \neq i} \frac{\pi_j}{1-\pi_j} + \ldots + \pi_i \sum_{j_1, \ldots, j_{|m^*|-1}} \frac{\pi_{j_1}}{1-\pi_{j_1}} \cdot \frac{\pi_{j_1} \cdots \pi_{j_{|m^*|-1}}}{1-\pi_{j_1} - \ldots \pi_{j_{|m^*|-1}}}.\end{aligned} \tag{11}$$

In the sampling literature the above probability is referred to as an inclusion probability. Observe that for large p the inclusion probability is approximately proportional to π_i. When $|m^*|\pi_i < 1$ for all i, one can use unequal probabilities (UP) sampling techniques, e.g Systematic Sampling proposed in [13], to have $P(i \in m^*) = |m|\pi_i$. For further examples of UP sampling methods see [14]. To illustrate the issue consider a simple example. Let $p = 3$, $|m^*| = 2$, $\pi_1 = 0.4$ and $\pi_2 = \pi_3 = 0.3$. It is easy to verify that in this case inclusion probabilities calculated from (11) are equal 0.74, 0.62 and 0.62, respectively, whereas the desired values are 0.8, 0.6 and 0.6.

In the WRSM procedure variables whose individual influence on response is more significant, have larger probability of being chosen to any of the random subspaces. Since in WRSM the relevant variables are more likely to be selected, we can limit the number of repetitions B in the main loop and reduce the computational cost of the procedure.

Let $\hat{\beta}_{i,\{i\}}$ be a least squares estimator based on univariate model $y \sim x_i$ and $T_{i,\{i\}}$ be the corresponding t-statistic. In WRSM we take $w_i^{(0)} = |T_{i,\{i\}}|$.

In the following theorem we determine asymptotic final scores assigned by the above procedure. Let $\mathcal{M}_{|m|}$ be the family of all subsets $\{i_1, \ldots, i_{|m|}\}$ of $\{1, \ldots, p\}$ (models) of size $|m|$ and $|\mathcal{M}_{|m|}| = \binom{p}{|m|}$ be its cardinality. Analogously let $\mathcal{M}_{i,|m|}$ be the family of all subsets of size $|m|$ containing variable i and note that $|\mathcal{M}_{i,|m|}| = \binom{p-1}{|m|-1}$. Let P^* by a resampling measure on $\mathcal{M}_{|m|}$ determined by point 4 of the algorithm. Thus a probability of choosing model m is given by

$$P^*(m) = \sum_{S(j_1,\ldots,j_{|m|})} \pi_{j_1} \frac{\pi_{j_2}}{1-\pi_{j_1}} \cdots \frac{\pi_{j_{|m|}}}{1-\pi_{j_1}-\ldots-\pi_{j_{|m|-1}}},$$

where $S(j_1, \ldots, j_{|m|})$ is a set of all permutations of indices $\{j_1, \ldots, j_{|m|}\}$. The expected value with respect to this distribution will be denoted by $\mathbf{E}^*$. In the case of deterministic $\mathbf{X}$ let

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

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 (see [1]) that under its assumptions in both cases $(n - |m|)^{-1}T^2_{i,m} \xrightarrow{a.s.} t_{i,m}$. Thus $t_{i,m}$ stands for asymptotic weight in RSM scheme. We state the result for WRSM procedure in the case when the number of predictors p is fixed and the initial weights $w_i^{(0)}$ are deterministic.

Theorem 2 *Let $(w_1^{(0)}, \ldots, w_p^{(0)})'$ be a deterministic vector. In the case of deterministic $\mathbf{X}$ assume that $\lambda(m)$ 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 $(\mathbf{Y}^{(n)}, \mathbf{X}^{(n)})_{n=1}^{\infty}$*

$$FS_i^* \xrightarrow{P^*} AFS_i := \frac{\sum_{m\in\mathcal{M}_{i,|m|}} t_{i,m}P^*(m)}{\sum_{m\in\mathcal{M}_{i,|m|}} P^*(m)}, \quad \text{as} \quad n \to \infty.$$

Thus FS_i^* is asymptotically equivalent to AFS_i, which is a weighted average of mean squared error of prediction $MSEP$, when the variable i is omitted from model m. The average is taken over all models m containing this variable. Note that the limiting value AFS_i equals conditional expectation $\mathbf{E}^*(t_{i,M}|i \in M)$, where $M = m$ is a random subset chosen by the procedure. Observe that in the original RSM we have $P^*(m) = 1/|\mathcal{M}_{|m|}|$ and then

$$AFS_i := \frac{1}{|\mathcal{M}_{i,|m|}|} \sum_{m \in \mathcal{M}_{i,|m|}} t_{i,m}.$$

Now we will discuss screening random subspace method (SRSM).

SRSM Algorithm

1. Input: observed data $(\mathbf{Y}, \mathbf{X})$, number of subset draws B, size of the subspace $|m| < \min(p, n)$.
2. For each variable i fit univariate model $y \sim x_i$ and compute weight of i-th variable $w_i^{(0)}$.
3. Let $\mathcal{M}_{screen} = \{i : w_i^{(0)} > \text{median}_{1 \le k \le p}(w_k^{(0)})\}$. RSM procedure is performed on data $(\mathbf{Y}, \mathbf{X}_{\mathcal{M}_{screen}})$.
4. Output: Ordered list of variables $\{i_1, \dots, i_{\lfloor p/2 \rfloor}\}$.

In SRSM procedure the preliminary screening based on univariate models is performed. Variables corresponding to the smallest weights $w_i^{(0)}$ are discarded and the RSM is performed on the remaining variables. This step reduces data dimensionality. Here the choice of the median as the threshold in 3 is arbitrary, in general it may depend on preliminary knowledge of researcher. As in WRSM the number of repetitions B can be limited to reduce the computational cost. The choice of initial weights is $w_i^{(0)} = |T_{i,\{i\}}|$.

The following example shows a similar screening procedure which is based on thresholding of absolute values of t-statistics. We give the formal justification of such procedure under conditions given below. Define $\mathcal{M}_{screen} = \{i : |\hat{\beta}_{i,\{i\}}| > r_n/n\}$, where r_n is a threshold sequence such that $r_n/n \to r < \min_{i \in t} |\beta_i|$. Let $\gamma_i = \lim_{n \to \infty} n^{-1} \mathbf{X}_i' \mathbf{X}_t \boldsymbol{\beta}_t$, for $i \notin t$.

Proposition 6 *Assume that columns of* $\mathbf{X}$ *are standardised, i.e. their sample means are zero and* $n^{-1}||\mathbf{X}_i||^2 = 1$ *for all* i*. Assume also that* σ *is known,* $n^{-1}\mathbf{X}_{k_1}'\mathbf{X}_{k_2} \to 0$*, for all pairs of relevant variables* $k_1, k_2 \in t$ *and* $\log(p) = o(n)$*. Then*

$$P(\mathcal{M}_{screen} \supset t) \to 1. \tag{12}$$

If $\max_{i \notin t} |\gamma_i| < r$ *we have*

$$P(\mathcal{M}_{screen} = t) \to 1. \tag{13}$$

The proof of the above Proposition is relegated to the Appendix. Note that under assumptions of Proposition 6 ordering of variables with respect to $|T_{i,\{i\}}|$ is equivalent to ordering with respect to $|\hat{\beta}_{i,\{i\}}| = |n^{-1}\mathbf{X}_i'\mathbf{Y}|$. Convergence in (12) indicates that with probability tending to one the true model t will be contained in the set of variables retained after the screening procedure. It follows from (13) that when dependence between spurious and relevant variables is not very strong, the true model t will be identified with probability tending to one, even when the number of all potential variables is large. In practise it is difficult to apply the above procedure since the proper choice of the threshold sequence r_n depends on an unknown parameter $\boldsymbol{\beta}_t$.

3 Model selection procedures

We briefly describe model selection procedure based on the RSM. 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 $(\mathbf{Y}^{\text{test}}, \mathbf{X}^{\text{test}})$ containing n_{test} observations be a test set. The following two-stage model selection procedure is performed.
Step 1. RSM procedure is performed on set $(\mathbf{Y}^t, \mathbf{X}^t)$. The covariates $\{1, \ldots, p\}$ are ordered with respect to RSM final scores

$$FS_{i_1} \geq FS_{i_2} \geq \ldots \geq FS_{i_p}.$$

Step 2. From the nested family of models

$$\mathcal{M}_{nested} = \{\{i_1\}, \{i_1, i_2\}, \ldots, \{i_1, i_2, \ldots, i_{\min(p,n)-1}\}\}$$

we select model $m_{\text{opt}} = \{i_1, \ldots, i_{|m_{\text{opt}}|}\}$ for which the prediction error $n_v^{-1}||\mathbf{Y}^v - \mathbf{X}^v\hat{\boldsymbol{\beta}}_{m_{\text{opt}}}||^2$ is minimal. Here, $\hat{\boldsymbol{\beta}}_{m_{\text{opt}}}$ is a least squares estimator based on model m_{opt} computed on training data. The analogous model selection procedure is performed for WRSM and SRSM using in step 1 the ordering given by the respective procedure.

The score FS_i is a variable importance measure which shows the significance of the i-th variable and describes its predictive power. In the first step we obtain a ranking of variables, showing what is the contribution of each of them in explaining the response. It follows from the properties of QR decomposition that in the second step it suffices to fit only one model based on $\min(n, p) - 1$ variables sorted according to ranks of final scores. If only variable importance estimation is of interest there is no need to split data into training and validation sets– the RSM is performed based on all n observations.

As benchmarks we also consider two other methods. The first, is the lasso method proposed in [3]. For this method, the estimator is defined by

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

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. $n_v^{-1}||\mathbf{Y}^v - \mathbf{X}^v\hat{\boldsymbol{\beta}}_{\text{lasso}}(\alpha)||^2$ or by cross-validation. We use the first option in our numerical experiments in order to make a comparison with the RSM more objective.

As the second benchmark, the univariate approach is considered. In this method informativeness and prediction strength of each feature is evaluated individually. Here, for each variable $i \in \{1, \ldots, p\}$ 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.

4 Numerical Experiments

In this section we study the performance of the proposed methods as prediction tools. We compare original RSM with WRSM and SRSM proposed here. We also used a hybrid method WSRSM in which in the first step screening is performed and then the WRSM is applied to remaining variables. As benchmarks we used the lasso and the univariate method. 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) $t = (2k + 7 : k = 3, \ldots, 12)$, $\boldsymbol{\beta}_t = (1, \ldots, 1)'$,

(M2) $t = (k^2 : k = 1, \ldots, 5)$, $\boldsymbol{\beta}_t = (1, 1, 1, 1, 1)'$,

(M3) $t = (1, \ldots, 5, 11, \ldots, 15, 21, \ldots, 25)$,
$\boldsymbol{\beta}_t = (2.5, \ldots, 2.5, 1.5, \ldots, 1.5, 0.5, \ldots, 0.5)'$.

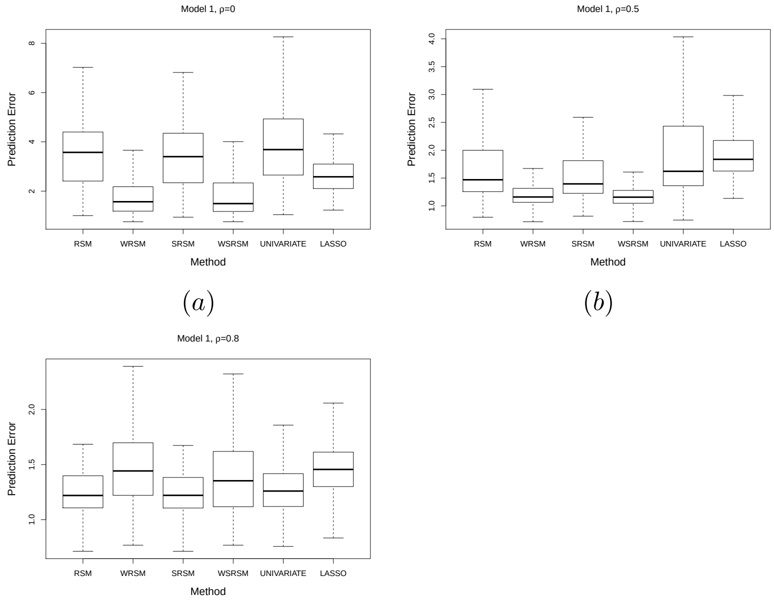
Number of potential regressors is $p = 1000$, and number of observations is $n = 200$. The rows of $\mathbf{X}$ were generated independently from the standard normal p-dimensional distribution with zero mean and the covariance matrix $\Sigma_{\mathbf{x}} = (\rho_{ij}) = \rho^{|i-j|}$. Three values of $\rho = 0, 0.5, 0.8$ were considered. The outcome is $\mathbf{Y} = \mathbf{X}_t \boldsymbol{\beta}_t + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ has zero-mean normal distribution with covariance matrix $\sigma^2 \mathbf{I}$ and $\sigma^2 = 1$ (for models M1 and M2) and $\sigma^2 = 1.5$ (for model M3). Models M1 and M2 were used in [15] whereas model M3 is model 7 in [16]. Observe that for models M1 and M2 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 $L = 500$ 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 $n_v/2 = 100$ observations each and final model m_{opt} is selected as described in Section 3.

For all methods the prediction strength of the selected model is assessed by prediction error on independent test set using the average error

$$n_{\text{test}}^{-1} ||\mathbf{Y}^{\text{test}} - \mathbf{X}^{\text{test}} \hat{\boldsymbol{\beta}}_{m_{\text{opt}}}||^2$$

with $\hat{\boldsymbol{\beta}}_{m_{\text{opt}}}$ being an estimator based on model m_{opt} computed on training data. For the RSM we considered $B = 5000$ choices of a random subspace consisting of $|m| = \min(n_t, p)/2 = 50$ attributes.

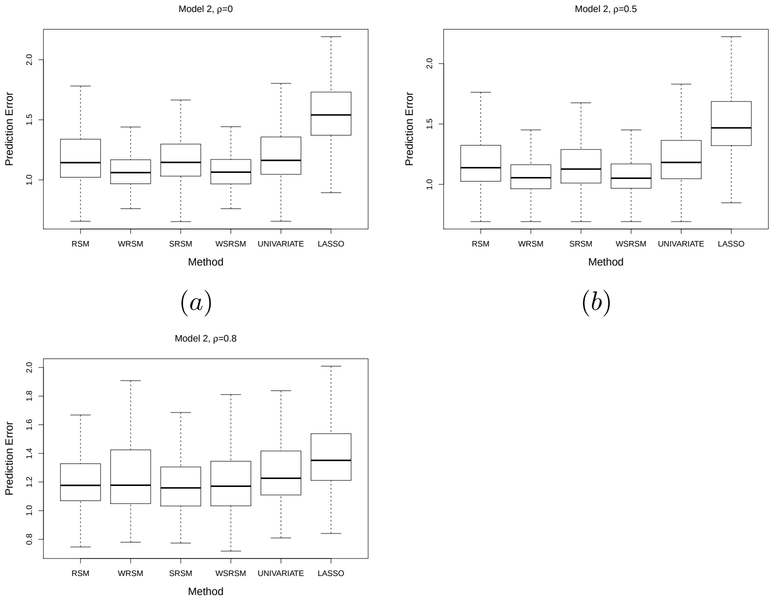
Figures 2, 3, 4 present prediction errors for models (M1), (M2) and (M3). It is seen that RSM works better than the lasso for model (M1) and (M2) when the dependence is moderate ($\rho = 0.5$) or strong ($\rho = 0.8$). In the case of model (M3) lasso outperforms RSM. Using weighting in RSM improves the results for models (M1) and (M2) when the dependence is not very strong ($\rho \leq 0.5$). In the case of model (M3) WRSM outperforms RSM for all dependence structures. For model (M3), where the lasso outperforms RSM, it is in its turn outperformed by WRSM. It is interesting that screening (SRSM) does not improve the results of RSM for $M = 1000$ and WSRSM behaves comparably to WRSM. However, it should be pointed out that using WRSM and SRSM we can substantially reduce the number of repetitions B. Figure 5 presents the means of prediction errors with respect to B in the case of model (M3). In particular, figure 5 (a) indicates that the mean of prediction error for WRSM with $B = 50$ is smaller than the one for RSM with $B = 1000$.



(a) (b)

(c)

FIGURE 2: Prediction errors for model M1 with $M = 1000$ and $n = 200$ based on $L = 500$ simulation trials. Figure (a) corresponds to $\rho = 0$, figure (b) to $\rho = 0.5$ and (c) to $\rho = 0.8$.



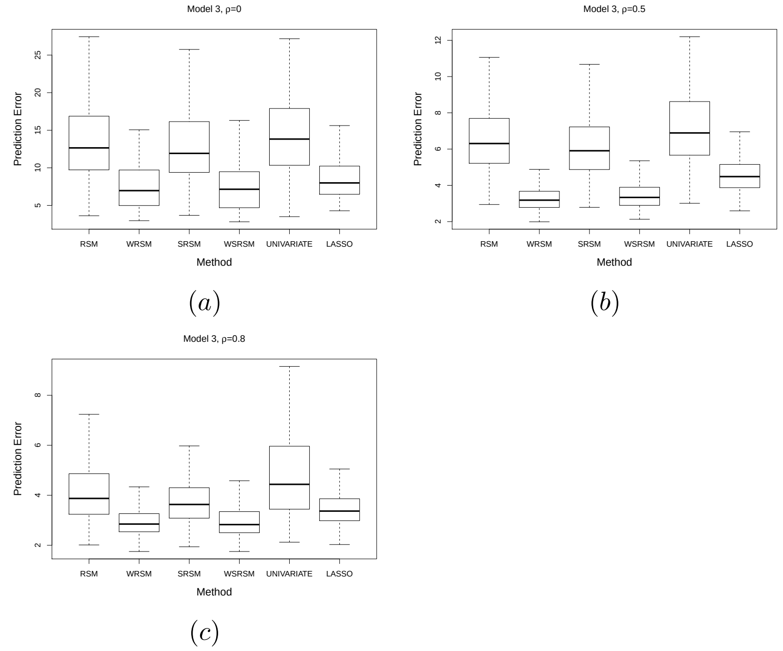
(a) (b)

(c)

FIGURE 3: Prediction errors for model M2 with $M = 1000$ and $n = 200$ based on $L = 500$ simulation trials. Figure (a) corresponds to $\rho = 0$, figure (b) to $\rho = 0.5$ and (c) to $\rho = 0.8$.

5 Acknowledgements

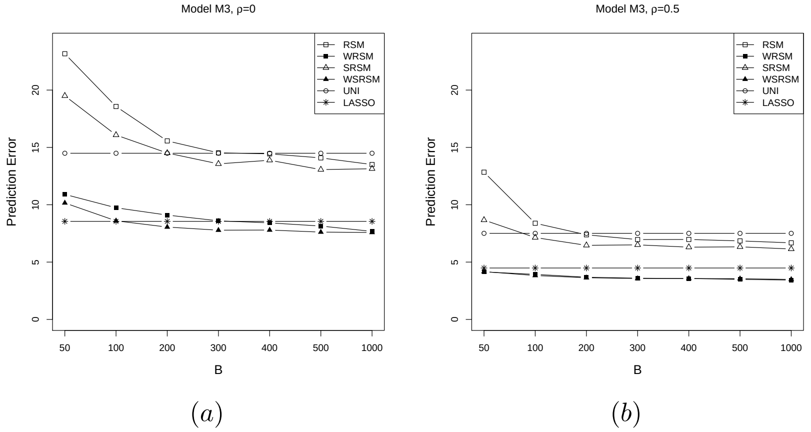
Research of the second author was partially supported by Project *'Information technologies: research and their interdisciplinary applications'* POKL.04.01.01-00-051/10-00. We appreciate the comments of anonymous referee which helped to improve the original version of the manuscript.



(a) (b)

(c)

FIGURE 4: Prediction errors for model M3 with $M = 1000$ and $n = 200$ based on $L = 500$ simulation trials. Figure (a) corresponds to $\rho = 0$, figure (b) to $\rho = 0.5$ and (c) to $\rho = 0.8$.



(a) (b)

FIGURE 5: Means of prediction errors with respect to B for model M3 with $M = 1000$ and $n = 200$ based on $L = 500$ simulation trials. Figure (a) corresponds to $\rho = 0$ and figure (b) to $\rho = 0.5$.

A Proofs

A.1 Proof of Theorem 2

First note that

$$\mathbf{E}^* \frac{T_{i,m^*}^2}{n - |m|} = \sum_{m \in \mathcal{M}_{i,|m|}} \frac{T_{i,m}^2}{n - |m|} P^*(m) \tag{14}$$

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

$$\mathrm{Var}^* \frac{T^2_{i,m^*}}{n-|m|} = \sum_{m\in\mathcal{M}_{i,|m|}} \frac{T^4_{i,m}}{(n-|m|)^2} P^*(m) - \left(\sum_{m\in\mathcal{M}_{i,|m|}} \frac{T^2_{i,m}}{n-|m|} P^*(m)\right)^2 \to$$
$$\sum_{m\in\mathcal{M}_{i,|m|}} t^2_{i,m} P^*(m) - \left(\sum_{m\in\mathcal{M}_{i,|m|}} t_{i,m} P^*(m)\right)^2 < \infty, \quad \text{as} \quad n\to\infty. \tag{15}$$

Using (14), (15) and Markov's inequality we have that

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

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

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

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

A.2 Proof of Proposition 2

Matrix W as a positive definite matrix can be decomposed as $W = W^{1/2}W^{1/2}$, where $W^{1/2} = US^{1/2}U'$, U is an orthogonal matrix and S is a diagonal matrix with positive diagonal. Let D_m be $(|t\cup m|)\times|m|$ matrix such that $\mathbf{X}_m = \mathbf{X}_{t\cup m}D_m$. We can write

$$\begin{aligned}
&n^{-1}||\mathbf{X}\boldsymbol{\beta} - P_m\mathbf{X}\boldsymbol{\beta}||^2 = \\
&n^{-1}\boldsymbol{\beta}'_t[\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m} - \mathbf{X}'_{t\cup m}\mathbf{X}_m(\mathbf{X}'_m\mathbf{X}_m)^{-1}\mathbf{X}'_m\mathbf{X}_{t\cup m}]\boldsymbol{\beta}_t = \\
&n^{-1}\boldsymbol{\beta}'_t[\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m} - \mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m}D_m(D'_m\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m}D_m)^{-1}D'_m\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m}]\boldsymbol{\beta}_t,
\end{aligned}$$

which converges to

$$\begin{aligned}
&\lambda(m) = \boldsymbol{\beta}'_t[W - WD_m(D'_mWD_m)^{-1}D'_mW]\boldsymbol{\beta}_t = \\
&(W^{1/2}\boldsymbol{\beta}_t)'[\mathbf{I} - W^{1/2}D_m[(W^{1/2}D_m)'(W^{1/2}D_m)]^{-1}D'_mW^{1/2}](W^{1/2}\boldsymbol{\beta}_t) = \\
&||(W^{1/2}\boldsymbol{\beta}_t) - H_m(W^{1/2}\boldsymbol{\beta}_t)||^2 > 0,
\end{aligned}$$

where H_m is a projection on the space spanned by columns of $W^{1/2}$. The last inequality follows from the fact that the columns of $W^{1/2}$ are linearly independent and model m does not contain at least one significant variable.

A.3 Proof of Proposition 3

Equality in (9) follows from

$$KL(\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}, \mathbf{f}_{\mathbf{X}_m\boldsymbol{\beta}_m}) = \int_{-\infty}^{+\infty} \mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}(s) \log \frac{\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t}(s)}{\mathbf{f}_{\mathbf{X}_m\boldsymbol{\beta}_m}(s)} ds =$$

$$\frac{2(\mathbf{X}_t\boldsymbol{\beta}_t)'(\mathbf{X}_t\boldsymbol{\beta}_t-\mathbf{X}_m\boldsymbol{\beta}_m)}{2\sigma^2}+\frac{(\mathbf{X}_m\boldsymbol{\beta}_m)'(\mathbf{X}_m\boldsymbol{\beta}_m)-(\mathbf{X}_t\boldsymbol{\beta}_t)'(\mathbf{X}_t\boldsymbol{\beta}_t)}{2\sigma^2}= \frac{||\mathbf{X}_t\boldsymbol{\beta}_t-\mathbf{X}_m\boldsymbol{\beta}_m||^2}{2\sigma^2}.$$

Equality (10) simply follows from

$$\inf_{\mathbf{X}_m\boldsymbol{\gamma}\in sp_{\mathbf{X}}(m)} KL(\mathbf{f}_{\mathbf{X}_t\boldsymbol{\beta}_t},\mathbf{f}_{\mathbf{X}_m\boldsymbol{\gamma}})=\inf_{\mathbf{X}_m\boldsymbol{\gamma}\in sp_{\mathbf{X}}(m)}\frac{||\mathbf{X}_t\boldsymbol{\beta}_t-\mathbf{X}_m\boldsymbol{\gamma}||^2}{2\sigma^2}= \frac{||\mathbf{X}_t\boldsymbol{\beta}_t-P_m\mathbf{X}_t\boldsymbol{\beta}_t||^2}{2\sigma^2}=\frac{\lambda_n(m)}{2\sigma^2}.$$

A.4 Proof of Proposition 4

The following equalities hold

$$\begin{aligned}&\lambda_n(m)=||\mathbf{X}_t\boldsymbol{\beta}_t-P_m\mathbf{X}_t\boldsymbol{\beta}_t||^2=\\&||\mathbf{X}_{t\setminus m}\boldsymbol{\beta}_{t\setminus m}-P_m\mathbf{X}_{t\setminus m}\boldsymbol{\beta}_{t\setminus m}+\mathbf{X}_{m\cap t}\boldsymbol{\beta}_{m\cap t}-P_m\mathbf{X}_{m\cap t}\boldsymbol{\beta}_{m\cap t}||^2=\\&||\mathbf{X}_{t\setminus m}\boldsymbol{\beta}_{t\setminus m}-P_m\mathbf{X}_{t\setminus m}\boldsymbol{\beta}_{t\setminus m}||^2=\boldsymbol{\beta}'_{t\setminus m}\mathbf{X}'_{t\setminus m}[\mathbf{I}-P_m]\mathbf{X}_{t\setminus m}\boldsymbol{\beta}_{t\setminus m}=\\&\boldsymbol{\beta}'_{t\setminus m}[\mathbf{X}'_{t\setminus m}\mathbf{X}_{t\setminus m}-\mathbf{X}'_{t\setminus m}\mathbf{X}_m(\mathbf{X}'_m\mathbf{X}_m)^{-1}\mathbf{X}'_m\mathbf{X}_{t\setminus m}]\boldsymbol{\beta}_{t\setminus m}.\end{aligned}$$

The second equality follows from the fact that P_m is linear.

A.5 Proof of Proposition 5

The following inequality holds

$$\begin{aligned}&\lambda_n(m)=||\mathbf{X}_t\boldsymbol{\beta}_t-P_m\mathbf{X}_t\boldsymbol{\beta}_t||^2=\inf_{\boldsymbol{\alpha}\in\mathbf{R}^{|m|}}||\mathbf{X}_{t\setminus m}\boldsymbol{\beta}_{t\setminus m}-\mathbf{X}_m\boldsymbol{\alpha}||^2=\\&\inf_{\boldsymbol{\alpha}\in\mathbf{R}^{|m|}}[(\boldsymbol{\beta}_{t\setminus m},\boldsymbol{\alpha})'\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m}(\boldsymbol{\beta}_{t\setminus m},\boldsymbol{\alpha})]\geq\lambda_{\min}(\mathbf{X}'_{t\cup m}\mathbf{X}_{t\cup m})||\boldsymbol{\beta}_{t\setminus m}||^2.\end{aligned}$$

A.6 Proof of Proposition 6

In view of the assumptions and the fact that $n^{-1}\mathbf{X}'_i\boldsymbol{\varepsilon}\sim N(0,\sigma^2/n)$ we have

$$\hat{\beta}_{i,\{i\}}=n^{-1}\mathbf{X}'_i\mathbf{Y}=n^{-1}||\mathbf{X}_i||^2\beta_i+n^{-1}\mathbf{X}'_i\mathbf{X}_{t\setminus i}\boldsymbol{\beta}_{t\setminus i}+n^{-1}\mathbf{X}'_i\boldsymbol{\varepsilon}\to\beta_i, \tag{16}$$

for relevant variable $i\in t$. For spurious variable $i\notin t$ we have

$$\hat{\beta}_{i,\{i\}}=n^{-1}\mathbf{X}'_i\mathbf{Y}=n^{-1}\mathbf{X}'_i\mathbf{X}_t\boldsymbol{\beta}_t+n^{-1}\mathbf{X}'_i\boldsymbol{\varepsilon}\to\gamma_i.$$

The convergence in (12) follows from (16). In order to show (13) we have to prove that $P(\max_{i\notin t}|\hat{\beta}_{i,\{i\}}|>r_n/n)\to 0$. The following inequalities hold

$$\begin{aligned}&P(\max_{i\notin t}|\hat{\beta}_{i,\{i\}}|>r_n/n)=P(\max_{i\notin t}|\mathbf{X}'_i\mathbf{Y}|>r_n)\leq(p-|t|)\max_{i\notin t}P(|\mathbf{X}'_i\mathbf{Y}|>r_n)\leq\\&(p-|t|)[\max_{i\notin t}P(\mathbf{X}'_i\mathbf{Y}>r_n)+\max_{i\notin t}P(\mathbf{X}'_i\mathbf{Y}<-r_n)].\end{aligned} \tag{17}$$

Now, using the assumption $\max_{i\notin t}|\gamma_i|<r$ and Mill's inequality (see in [17]), the first probability in (17) can be bounded from above by

$$p\max_{i\notin t}P(\mathbf{X}'_i\mathbf{X}_t\boldsymbol{\beta}_t+\mathbf{X}'_i\boldsymbol{\varepsilon}>r_n)=$$

$$p \max_{i \notin t} P(n^{-1/2}\mathbf{X}_i'\boldsymbol{\varepsilon} > n^{1/2}(r_n/n - \mathbf{X}_i'\mathbf{X}_t\boldsymbol{\beta}_t/n)) \leq$$
$$p \max_{i \notin t} \frac{1}{n^{1/2}(r_n/n - \mathbf{X}_i'\mathbf{X}_t\boldsymbol{\beta}_t/n)} \exp(-n(r_n/n - \mathbf{X}_i'\mathbf{X}_t\boldsymbol{\beta}_t/n)^2/2) \to 0,$$

under assumption $\log(p)/n \to 0$. The second probability in (17) is treated analogously.

References

[1] Mielniczuk, J., Teisseyre, P.: Using random subspace method for prediction and variable importance assesment in regression. Computational Statistics and Data Analysis (2012) in press, 10.1016/j.csda.2012.09.018.

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

[3] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B **58** (1996) 267–288

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

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

[6] Chen, J., Chen, Z.: Extended Bayesian Information Criteria for model selection with large model spaces. Biometrika **95** (2008) 759–771

[7] Frommlet, F., Ruhaltinger, F., Twaróg, B., Bogdan, M.: Modified versions of Bayesian information criterion for genome-wide association studies. Computational Statistics and Data Analysis **56** (2012) 1038–1051

[8] Zhang, P.: On the distributional properties of model selection criteria. Journal of the American Statistical Association **87** (1992) 732–737

[9] Shao, J.: Linear model selection by cross-validation. Journal of the American Statistical Association **88** (1993) 486–494

[10] Zheng, X., Loh, W.Y.: Consistent variable selection in linear models. Journal of the American Statistical Association **90**(429) (1995) 151–156

[11] Casella, G., Giron, J., Martinez, M., Moreno, E.: Consistency of Bayesian procedures for variable selection. Annals of Statistics **37** (2009) 1207–1228

[12] Gentle, J.E.: Matrix Algebra: Theory, Computations, and Applications in Statistics. Springer, New York (2007)

[13] Madow, W.G.: On the theory of systematic sampling, II. Annals of Mathematical Statistics **20** (1949) 334–354

[14] Tillé, Y.: Sampling Algorithms. Springer, New York (2006)

[15] Zheng, X., Loh, W.Y.: A consistent variable selection criterion for linear models with high-dimensional covariates. Statistica Sinica **7** (1997) 311–325

[16] Huang, J., Ma, S., Zhang, C.H.: Adaptive lasso for high-dimensional regression models. Statistica Sinica **18** (2008) 1603–1618

[17] Birnbaum, Z.W.: An inequality for Mill's ratio. Annals of Mathematical Statistics **13**(2) (1942) 245–246