A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 286, 2013

Mariusz Kubus[*]

FEATURE SELECTION IN HIGH DIMENSIONAL REGRESSION PROBLEM

Abstract. There are three main approaches to feature selection problem considered in statistical and machine learning literature: filters, wrappers and embedded methods. Filters evaluate and exclude some variables before learning a model. Wrappers use learning algorithm for evaluation of the feature subsets and involve search techniques in the feature subset space. Embedded methods use feature selection as an integral part of learning algorithm. When features outnumber examples, filters or embedded methods are recommended. The goal of this paper is to compare popular filters and embedded methods in high dimensional problem. In the simulation study, redundant variables will be included in the artificially generated data.

Key words: feature selection, filters, embedded methods, high dimension.

I. INTRODUCTION

Generally there are two ways of dealing with high dimension in predictive modelling. The first approach uses transformation of original data matrix $[\boldsymbol{X}, Y]$ as a pre-processing step to reduce a dimension (i.e. principal components). The second approach – the one we focus on in this paper – selects original features using various criteria. The methods of feature selection are currently classified into three groups: filters, wrappers and embedded methods (see i.e. Blum and Langley (1997); Guyon and Elisseeff (2003)). All of them perform a search in the space of all possible subsets of variables. The third group differs considerably because the search is an integral part of a learning algorithm. Wrappers and filters have much in common. Both can be seen as search task where one must choose the function of criterion, the search strategy and the stopping criterion to obtain an optimal subset of features. The difference between wrappers and filters is that the first ones use model evaluation to assess a subset of variables (i.e. stepwise regression), while the letter ones eliminate variables in pre-processing step and are independent of the learning algorithm. Various functions of criterion for filters were proposed in the literature (for a survey see i.e. Nowak (1984); Grabiński et al. (1982); Duch (2006)). In case of

[*] Ph.D., Department of Mathematics and Applied Computer Science, Opole University of Technology.

wrappers, prediction error estimated by cross-validation or information criteria are usually used for evaluation purposes. The most commonly used search strategies are heuristic (i.e. greedy search, best first) or random (genetic algorithm or simulated annealing). For more details on search strategies see (Reunanen 2006).

The special case when the number of variables p exceeds the number of examples N is especially treated in the literature (i.e. Hastie et al. (2009)). It is typical in computational biology in the analysis of microarrays of gene expression and some recently proposed feature selection methods are specifically addressing this problem (i.e. Zou and Hastie (2005); Meinshausen (2007); Paul et al. (2008)). In economic sciences one can deal with the case $p > N$ having introduced to the model specification the interaction terms or functions of original variables. Predictive modelling in the case when $p > N$ is difficult due to: lack of degrees of freedom, overfitting problem and instability of estimated coefficients. In practice, the problems with collinearity also appear.

The goal of this paper is to test and compare some filters and embedded methods (for regression) in high dimension ($p > N$). We found it interesting to compare such sophisticated methods like relaxed LASSO or regression trees with simple filters (well known from Polish handbooks of econometrics). The mix of filters and wrappers will also be investigated. The artificial data used in simulation study will have included redundant variables.

II. SOME FILTER METHODS

Filter methods are widely used in the situation when the number of variables exceeds the number of examples. They usually perform univariate scoring which means that every predictor is evaluated in turn independently of others. The problem of feature selection is then simplified. Nevertheless, it works quite well in practice and provides considerable benefits with regards to computation costs. Moreover, most advanced techniques can lead to overfitting (Guyon 2008). All filters described in this section perform univariate feature ranking which in various ways uses the correlation measures.

The most basic filter excludes variables which are not significantly correlated with the response. Additionally, one can apply the second feature selection method in previously reduced space (i.e. stepwise regression which is commonly available in most statistical software). The second filter considered in this paper was proposed by Nowak (1997). It uses correlation analysis and its fundamental idea is to identify predictors highly correlated with response and at the same time, poorly correlated among themselves. In this way, information

about response is not doubled. In the first step predictors correlated with response below predefined threshold are excluded. Next two steps are performed as long as any predictor is available. The predictor which is most correlated with the response is chosen and subsequently predictors correlated with it above predefined threshold are excluded.

The next filter considered in this paper is in fact a mix of filter and wrapper approach (see i.e. Guyon 2008). Firstly, predictors are ordered according to a chosen measure $X_{j_1} \succ X_{j_2} \succ X_{j_3} \succ ... \succ X_{j_p}$ (i.e. absolute correlation with response) and then nested models are constructed (linear models in this paper). It means that successive models are learned with a use of subsets of variables respectively $\{j_1\}, \{j_1, j_2\}, \{j_1, j_2, j_3\}...$ Models are evaluated using prediction error estimated by 10-fold cross-validation and the best one is chosen. Such approach sufficiently decreases computational burden in comparison to wrapper. Moreover, Ng (1998) showed that this technique is less prone to overfitting than pure wrapper methods. Additionally, one standard error rule can be applied to obtain sparser model which yields prediction error not greater than one standard error above minimum obtained by cross-validation.

III. SOME EMBEDDED METHODS

The second recommended approach to feature selection in case when $p > N$ is a use of embedded methods. They are less computationally intensive and less prone to overfitting than wrappers. The examples are tree based models or regularized linear models. In this paper we focus on regression trees (see Gatnar 2001), LASSO (Tibshirani 1996) and its approximation solution LARS (Efron et al. 2004), and its modified version (relaxed LASSO) proposed by (Meinshausen 2007).

The LASSO estimates are defined by:

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \left(\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \cdot \sum_{j=1}^{p} \left| \beta_j \right| \right). \tag{1}$$

The parameter of regularization $\lambda \geq 0$ causes shrinkage of the estimators towards 0, and in practice some of them may be exactly 0 (what is equivalent to feature selection). The problem of tuning of the λ parameter is usually solved by cross-validation. Unfortunately, the regularization task (1) does not have a solution in the closed matrix form as in ridge regression, and quadratic

programming with linear constraints must be performed. Most commonly used are various approximation methods (i.e. incremental forward stagewise (Hastie et al. 2009), homotopy method (Osborne et al. 2000), LARS (Efron et al. 2004)).

The last mentioned method iteratively estimates the coefficients of the linear model. Predictors are standarized at the beginning and all coefficients are equal to zero. At every step, the model is fitted to the current residuum and variable most correlated with it joins the model. Geometrically one can imagine this process as shifting of a point which represents a fitted values $\hat{\mathbf{y}}$ towards the OLS solution. The direction of this shifting in the k-th step keeps equal angels with k predictors previously introduced into the model. Thus, algorithm needs to run no more than p steps. As a result, the family of nested models is obtained which differ with regards to the number of predictors. The last stage is to make decision how many variables the model should consist of. Usually prediction error estimated by cross-validation (alternatively with a one standard error rule) is used for that purpose.

Due to difficulty with controlling shrinkage and model selection with a use of only one parameter (Meinshausen 2007) proposed modified version of LASSO. The relaxed LASSO estimate is performed in two steps. Firstly, LARS algorithm is applied for feature selection and then second tuning parameter controls shrinkage for previously selected features. Meinshausen showed some advantages over ordinary LASSO in high dimensional problems. The number of selected features in relaxed LASSO is in general much smaller and it yields more accurate predictions for a high signal-to-noise ratios. Note however that the results obtained by (Meinshausen 2007) relate to orthogonal design.

Regression trees are a nonparametric and adaptive method. It means that the assumption of an analytic form of the model is not required and model is fitted to the data locally. The multidimensional feature space is recursively partitioned into disjoint regions and the response is estimated as a constant in each of them. The regions are defined by chosen variables and splitting points, and their borders are parallel to the axes. Variables are introduced into the model so that to minimize the variance in the regions. The resulted tree is usually pruned to obtain the trade-off between complexity (many nodes and variables introduced into a model) and the accuracy of prediction (for more details see (Gatnar 2001)).

IV. SIMULATION STUDY

The goal of the experiment is to compare filters and embedded methods presented above in the case when $p > N$. In the simulation, we used two models: linear (2) and linear with interactions (3):

$$y = \beta_0 + \sum_{j=1}^{5} \beta_j x_j + e\,, \tag{2}$$

$$y = \beta_0 + \sum_{j=1}^{5} \beta_j x_j + \beta_6 x_1 x_2 + \beta_7 x_3 x_4 + e\,. \tag{3}$$

Predictors were generated from univariate standarized normal distribution and there were no correlations between them. Gaussian noise $e \sim N(0,s)$, where $s = 0.1 \cdot sd(y)$, was added only to training sets. The sizes of the training sets and test sets were chosen to be 100 and 500 respectively. Furthermore, 200 irrelevant variables Z_j were introduced to the data sets. First 100 variables Z_j were independent. In the next 50 variables Z_j multicollinearity were introduced so that every fifth variable was a linear combination of the previous four:

$$Z_{5+k*5} = \alpha_{k1} Z_{1+k*5} + \alpha_{k2} Z_{2+k*5} + \alpha_{k3} Z_{3+k*5} + \alpha_{k4} Z_{4+k*5} + e_j\,, \tag{4}$$

for $k \in \{20,...,29\}$, where $e_j \sim N(0, s_j)$ and $s_j = h \cdot sd(Z_j)$. The level of the noise h was sampled from the set $\{0.1,0.2,0.3\}$. The last 50 variables Z_j were pairwise correlated according to the formula:

$$Z_{150+k*2+2} = \gamma_k Z_{150+k*2+1} + e_j\,, \tag{5}$$

for $k \in \{0,1,2,...,24\}$, where e_j was set as above and γ_k was sampled from the set $\{0.5,\ 0.75,\ 1,\ 1.25,\ 1.5\}$. All coefficients as well as realizations of the variables (despite dependent variables in formulas (4) and (5)) were generated from univariate standardized normal distribution.

The filters we examined include: Spearman correlation of ranks (S), also followed by stepwise regression procedure (S+fs), filter which follows Nowak's scenario (N), also followed by stepwise regression procedure (N+fs) and the combination of filter and wrapper (F-W). We chose the Spearman correlation of ranks because the test of significance is free of assumption about normal distribution and it is able to capture any monotonic dependence. This can be of significant importance in models with interactions. The stepwise regression was applied in the version of forward selection. The filter-wrapper combination was applied with one standard error rule.

In all embedded methods examined we use cross-validation for model selection. In case of LASSO one standard error rule is applied. Note that we experimented with AIC criterion but it did not yield promising results.

Regression trees were applied with cost-complexity pruning. All simulations were conducted using R codes with a use of packages: lars, relaxo and rpart.

For the linear model (2), the results of the simulations are summarized in Figure 1. The medians of prediction errors (estimated from test sets) are the lowest for relaxed LASSO and second lowest for LASSO. However, the second mentioned method does not detect the irrelevant variables effectively. The median of the number of irrelevant variables for relaxed LASSO is equal to zero and it is lower than medians obtained by filters. On the other hand there are many outliers in the right tail of the empirical distribution (18%). Note that additional stepwise forward selection procedure which follows filters S and N considerably improves the capability of detecting of irrelevant variables while the prediction errors are comparable.

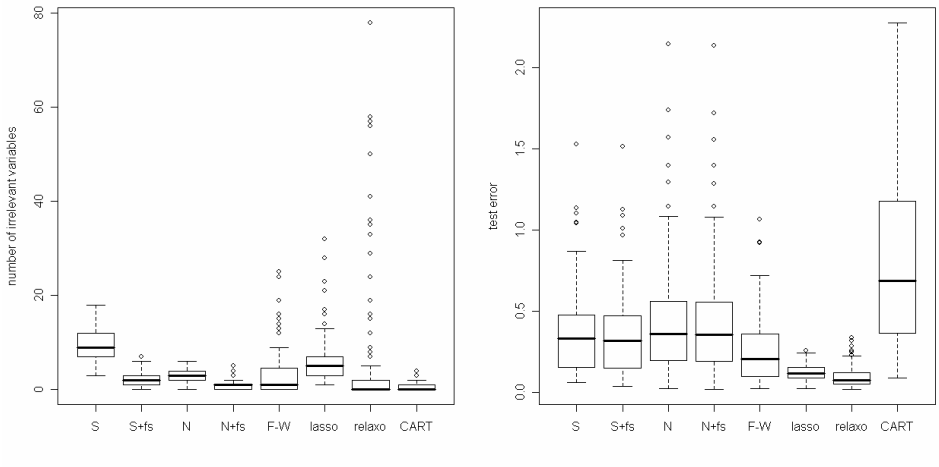

Fig. 1. Results over 100 simulations for linear model (2).
Source: own computations.

The results obtained for the model with interactions (3) are summarized in Figure 2. Here the differences in prediction errors are not so distinct. This might result from incorrect model specification (it is not known at the beginning of the analysis and the dimension is too high to introduce the interaction terms and functions of the original variables). As in case of linear model (2) the results point towards applying stepwise forward selection after filters S or N. The median of the number of irrelevant variables added to the models is equal zero for F-W and relaxed LASSO, but F-W is slightly more stable (lower standard deviation 3.1 when relaxed LASSO yields 8).

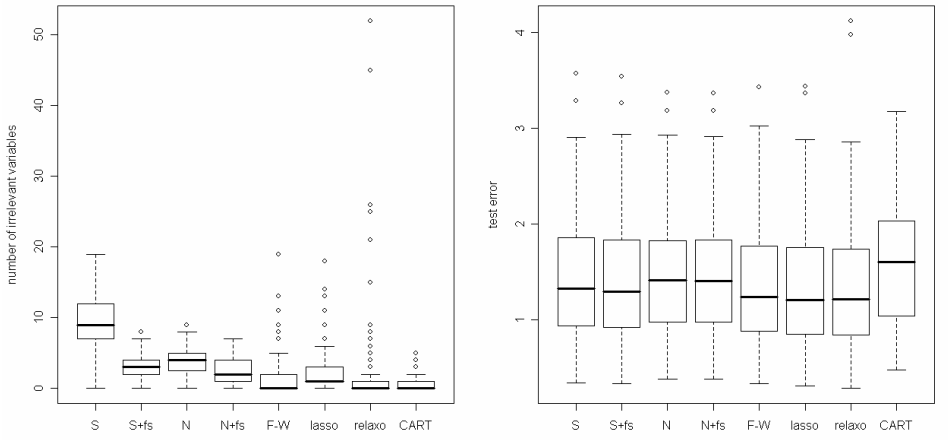

Fig. 2. Results over 100 simulations for model with interactions (3).
Source: own computations.

V. CONCLUSIONS

The popular filters and embedded methods were compared with regards to prediction error and capability to detect the noisy variables in the case when $p > N$. Especially worth of paying closer attention to is the relaxed LASSO which outperformed other methods when the model was linear. In the presence of interactions it works as well as filter-wrapper procedure and these two are the most recommended methods. Note that instability is a significant drawback of relaxed LASSO. Sometimes it includes into the model a great number of irrelevant variables. The popular filters are not as effective as relaxed LASSO. Nowak's procedure turned out to be too radical in removing irrelevant variables and it usually yields higher prediction errors than other methods. CART is most radical in discarding less informative features and it clearly leads to underfitting.

REFERENCES

Blum A.L., Langley P. (1997), Selection of relevant features and examples in machine learning. „*Artificial Intelligence*", vol. 97 no. 1-2, p. 245-271.

Duch W. (2006), Filter methods. [in:] Guyon I., Gunn S., Nikravesh M., Zadeh L. (Eds.), *Feature Extraction: Foundations and Applications*. Springer, New York.

Efron B., Hastie T., Johnstone I., Tibshirani R. (2004), Least Angle Regression. „*Annals of Statistics*" 32 (2): p. 407-499.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*. PWN, Warszawa.

Grabiński T., Wydymus S., Zeliaś A. (1982), *Metody doboru zmiennych w modelach ekonometrycznych*. PWN, Warszawa.

Guyon I. (2008), Practical Feature Selection: from Correlation to Causality. [in:] F. Fogelman-Soulie et al. (Eds.), *Mining Massive Data Sets for Security*, IOS Press.
Guyon I., Elisseeff A. (2003), An Introduction to Variable and Feature Selection. „*Journal of Machine Learning Research*" 3, p. 1157-1182.
Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning: Data Mining, Inferance, and Prediction.* 2nd edition, Springer, New York.
Meinshausen N. (2007), Lasso with relaxation, *Computational Statistics and Data Analysis* 52(1): p. 374-293.
Ng A.Y. (1998), On feature selection: learning with exponentially many irrelevant features as training examples, In *Proceedings of the 15th International Conference on Machine Learning*, p. 404-412, San Francisco, CA. Morgan Kaufmann.
Nowak E. (1984), *Problemy doboru zmiennych do modelu ekonometrycznego*. PWN, Warszawa.
Nowak E. (1997), *Zarys metod ekonometrii: zbiór zadań.* PWN Wyd.2, Warszawa.
Osborne M., Presnell B., Turlach B. (2000), A new approach to variable selection in least squares problems. „*IMA Journal of Numerical Analysis*" 20: p. 389-404.
Paul D., Bair E., Hastie T., Tibshirani R. (2008), "Pre-conditioning" for feature selection and regression in high-dimensional problems, *Annals of Statistics* 36(4): p. 1595-1618.
Reunanen J. (2006), Search Strategies, In I. Guyon, S. Gunn, M. Nikravesh, L. Zadeh (Eds.), *Feature Extraction: Foundations and Applications,* Springer, New York.
Tibshirani R. (1996), Regression shrinkage and selection via the lasso. „*J.Royal. Statist. Soc. B.*" 58: p. 267-288.
Zou H., Hastie T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B.* 67(2): p. 301-320.

Mariusz Kubus

SELEKCJA ZMIENNYCH DLA REGRESJI W PRZYPADKU DUŻEGO WYMIARU PRZESTRZENI CECH

Metody selekcji zmiennych dyskutowane obecnie w literaturze dzielone są na trzy główne podejścia: dobór zmiennych dokonywany przed etapem budowy modelu, przeszukiwanie przestrzeni cech i selekcja zmiennych na podstawie oceny jakości modelu oraz metody z wbudowanym mechanizmem selekcji zmiennych. W przypadku, gdy liczba zmiennych jest większa od liczby obserwacji rekomendowane są głównie podejścia pierwsze lub trzecie. Celem artykułu jest porównanie wybranych metod reprezentujących te podejścia w przypadku dużego wymiaru przestrzeni cech. W przeprowadzonych symulacjach, do sztucznie generowanych danych włączano zmienne skorelowane.