ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 285, 2013

Mariusz Kubus[*]

ON MODEL SELECTION IN SOME REGULARIZED LINEAR REGRESSION METHODS

Abstract. A dynamic development of various regularization formulas in linear models has been observed recently. Penalizing the values of coefficients affects decreasing of the variance (shrinking coefficients to zero) and feature selection (setting zero for some coefficients). Feature selection via regularized linear models is preferred over popular wrapper methods in high dimension due to less computational burden as well as due to the fact that it is less prone to overfitting. However, estimated coefficients (and as a result quality of the model) depend on tuning parameters. Using model selection criteria available in R implementation does not guarantee that optimal model will be chosen. Having done simulation study we propose to use EDC criterion as an alternative.

Key words: model selection, EDC, regularization, linear models, feature selection.

I. INTRODUCTION

The model selection task is strictly connected to feature selection. Both these tasks are equivalent when one does not take into account the interaction terms or additional variables which are the functions of input variables in linear models. The feature selection has recently seen plenty of interest and it has gained a special position in data mining. Discarding the irrelevant variables from the data improves the predictive accuracy of the models (overfitting avoidance). Identification of the most informative features gives some insight into data generation process and it can minimize the costs of data collection in the future. The methods of feature selection are currently classified into three groups: filters, wrappers and embedded methods (i.e. see Guyon at al. 2006). Regularized linear models which represent embedded methods approach are particularly recommended in high dimension. They are relatively fast and less prone to overfitting than other techniques. Unlike commonly applied univariate filters, regularized linear models can capture the joint impact of predictors on the response. However, the main problem with applying them is a setting the regularization parameters. It has the decisive impact on the quality of the models. In practice the

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

regularization parameters are set using model selection criteria. Unfortunately, using the criteria available in R implementation does not guarantee that optimal model will be chosen. The goal of this paper is to verify the usefulness of various model selection criteria in some regularized linear models.

II. REGULARIZED LINEAR REGRESSION

In regression problem one has given a vector of predictors $\boldsymbol{X} = (X_1,...,X_p)$ and real-valued response *Y*. Given *N* multidimensional observations

$$\{(\boldsymbol{x_1}, y_1),...,(\boldsymbol{x_N}, y_N) : \boldsymbol{x_i} \in \boldsymbol{X} = (X_1,...,X_p), y_i \in Y, i \in \{1,...,N\}\}$$

the goal is to estimate the function: : $y = f(\boldsymbol{x})$. In regularized linear models parameters are estimated by minimizing the sum of loss function (often squared loss) and penalty component:

$$\hat{\beta} = \arg\min_{\beta} \left(\sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + P(\boldsymbol{\lambda}, \boldsymbol{\beta}) \right). \qquad (1)$$

Various methods proposed in the literature differ mainly with regards to the penalty component $P(\boldsymbol{\lambda}, \boldsymbol{\beta})$. One of the most popular ones (the sum of absolute values of the coefficients, so called LASSO) was proposed by (Tibshirani 1996). Zou and Hastie (2005) proposed penalty component (*elastic net*) which is a compromise between ridge regression and LASSO:

$$P(\lambda, \alpha, \boldsymbol{\beta}) = \lambda \cdot \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1-\alpha)|\beta_j| \right). \qquad (2)$$

These methods perform shrinking of the coefficients to zero (what results with decreasing the variance) and feature selection (when some coefficients are equal zero). The amount of shrinkage and at the same time the quality of the model depend on tuning of the parameters which are set in practice with a use of model selection criteria. The estimation of the parameters in LASSO requires quadratic programming with linear constrains. Approximate solutions are more practical and more commonly used. Efron et al. (2004) proposed LARS algorithm which is very fast and probably the most popular algorithm currently available. The R implementation of the *elastic net* is also based on it.

In LARS, variables are successively included into the model and coefficients are iteratively updated. The criterion is based on correlation. The model is fitted to the current residuum at every step and algorithm requires no more steps than the number of predictors. As a result, the family of nested models which differ with the number of predictors is obtained. The last stage is to decide how many variables the model should consist of. It affects mainly the quality of the final model.

As mentioned in Introduction, the R implementation (packages: `lars` and `elasticnet`) offers C_p statistic or prediction error estimated by cross-validation for that purpose. As standard errors of estimations by cross-validation are also directly available, the application of one standard error rule (1SE rule) is possible. It means choosing the model with a lesser number of parameters whose error is no more than one standard error above the error of the best model. Such strategy is applied to avoid overfitting. As we will see in the experiment, implemented criteria do not always lead to optimal model. In the next section we list some alternatives for model selection.

III. MODEL SELECTION CRITERIA

Usually the quality of the model is defined as a capability to generalization which means the accurate prediction for future observations (out of training set). As the estimation of prediction error in training set is too optimistic there are generally two approaches proposed in the literature.

The first approach uses the correction for mentioned optimism. Some of the propositions are listed in (Maddala 2008, p.548) but we focus at currently more popular criteria which are expressed as a sum of two components: prediction error (often increasing function of this error) and the penalty for a number of parameters. As the error is estimated in the training set, models with a greater number of parameters yield more accurate (in-sample) predictions, but the penalty is higher. On the other hand, underestimated models yield greater errors although the penalty is low. Thus, the criterion is a compromise between goodness of fit and complexity of the models. The problem of model selection can also be described from the prediction error decomposition perspective or in other words via bias-variance tradeoff (see i.e. Hastie et al. 2009). Note that in this paper we have limited our interest to the linear models without interaction terms and variables being functions of input variables. In such cases k in formulas below denotes the number of original variables. Mentioned C_p statistic takes a form:

$$C_p(k) = \frac{RSS}{\hat{\sigma}^2} + 2(k+1) - N \,, \tag{3}$$

where: *RSS* is a residual sum of squares for model with *k* parameters, *N* is the number of observations. As the variance is estimated for the model with all input variables, the criterion cannot be applied in case when $p > N$. Commonly used for model selection are information criteria. They can be written in general form as:

$$Q(k) = N \cdot \log\left(\frac{RSS}{N}\right) + (k+1) \cdot P(N), \tag{4}$$

where *P*(*N*) is a function that decides about amount of penalty for the number of parameters. Several criteria proposed in the literature differ with a term *P*(*N*) (for a review see i.e. Kundu and Murali 1996). The most popular are AIC criterion with $P(N) = 2$ and BIC with $P(N) = \log(N)$. Bai et al. (1986) proposed to apply the penalty which would satisfy the conditions (EDC criterion):

$$\lim_{N\to\infty} \frac{P(N)}{N} = 0 \quad \text{and} \quad \lim_{N\to\infty} \frac{P(N)}{\log\log N} = \infty \,. \tag{5}$$

For example $P(N)$ can be a square root of the number of observations. Kundu and Murali (1996) showed empirically that this criterion very radically decreases the number of parameters. Note that BIC also satisfies the conditions (5). Hurvich and Tsai (1989) proposed the modified version of AIC, recommended especially in the case of small training samples or a great number of variables (see Burnham and Anderson (2002)):

$$AICc = AIC + \frac{2k(k+1)}{N-k-1} \,. \tag{6}$$

The second approach to model selection is a direct estimation of out of sample prediction error using a validation set (with known values of response). In practice, such validation set is usually not available. If data are huge one can divide them into training set, validation set for model selection, and test set for model assessment. In case of a small number of observations cross-validation or bootstraping is performed. The number of folds in cross-validation depends on the training set size and on how the performance of the learning method varies with that size. Breiman and Spector (1992) recommend 5 or 10-folds. In case of small training sets leave-one-out cross-validation is performed (every fold contains one observation), which has low bias but can have a high variance. Such approach demands constructing the model *N* times, thus one can apply a convenient approximation (for linear models) using generalized cross-validation criterion (Wahba 1980):

$$GCV(k) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{1 - \frac{k+1}{N}}\right)^2 . \tag{7}$$

It is worth mentioning that cross-validation performed for model selection cannot be treated as an assessment of the final model. For that purpose a different test set is necessary.

IV. EXPERIMENT

The goal of the performed experiments is to assess the usefulness of known model selection criteria which are not implemented in R packages `lars` and `elasticnet`. We take them into account from the perspective of three potential advantages. Firstly – the accuracy of the obtained models, secondly – the number of irrelevant variables introduced into the model and thirdly – the computational burden. The simulations are carried out according to the linear model:

$$y = \beta_0 + \sum_{j=1}^{10}\beta_j x_j + e . \tag{8}$$

The sizes of the train samples were chosen to be 200 in the first two experiments and 50 in the third. The test samples always contained 500 observations. There were also noisy variables added to the data – various numbers in the particular experiments. Multicollinearity between predictors was introduced according to the formula:

$$x_{5+k*5} = \alpha_{k1}x_{1+k*5} + \alpha_{k2}x_{2+k*5} + \alpha_{k3}x_{3+k*5} + \alpha_{k4}x_{4+k*5} + e_k , \tag{9}$$

for $k \in \{0,1,..., p/5-1\}$, where p denotes the number of predictors. Note that relevant variables are collinear as well as a noisy. All coefficients α_{ki}, β_j $(i = 1,...4, j = 0,...,10)$ as well as the realizations of variables (despite x_{5+k*5}

which are the linear combinations of another predictors) were generated from univariate standarized normal distribution. The noise in formulas (8) and (9) was normally distributed with expected values 0 and standard deviation being the standard deviation of dependent variable (y or x_{5+k*5} respectively) multiplied by randomly chosen $m \in \{0.1,0.2,0.3,0.4\}$.

Experiment 1

There were 10 noisy variables added to the data. Figure 1 depicts the numbers of irrelevant variables included to the models (left panels) and test errors (right panels) over 100 simulations. For both algorithms (LARS and *elastic net*) EDC includes fewest irrelevant variables to the models. At the same time test errors are comparable, although EDC yields occasionally high errors (three or four outliers). We confirmed it using post-hoc tests after Friedman's test (Nemenyi 1963). The lowest mean error was obtained using BIC. It was significantly lower than the one obtained by C$_p$ and did not differ significantly from the one obtained by cross-validation with 1SE rule (CV+1SE).

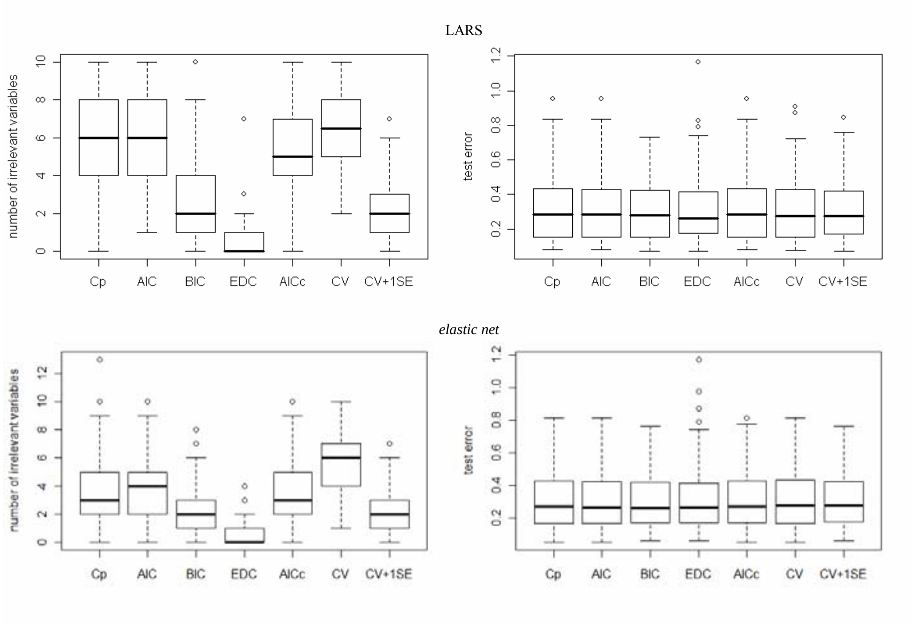

Figure 1. The comparison of seven model selection criteria in the case N=200 and p=20.
Source: own computations.

Experiment 2

There were 90 noisy variables added to the data. The results are presented in the similar way in Figure 2. Again we can conclude that EDC, BIC and CV+1SE outperform other criteria. When the number of variables has grown to 100 the differences became clearer. It can also be seen with regards to test errors. The post-hoc tests after Friedman's test showed that EDC does not differ significantly from BIC or CV+1SE and significantly outperforms other criteria. In *elastic net* for example, the absolute values of differences between average of the ranks (between EDC and other criteria) were: 2.15, 3.61, 0.32, 0.00, 2.55, 3.62, 0.82 respectively. The critical difference was 0.90 for significance level 0.05. At the same time EDC is the most radical criterion in discarding irrelevant variables.

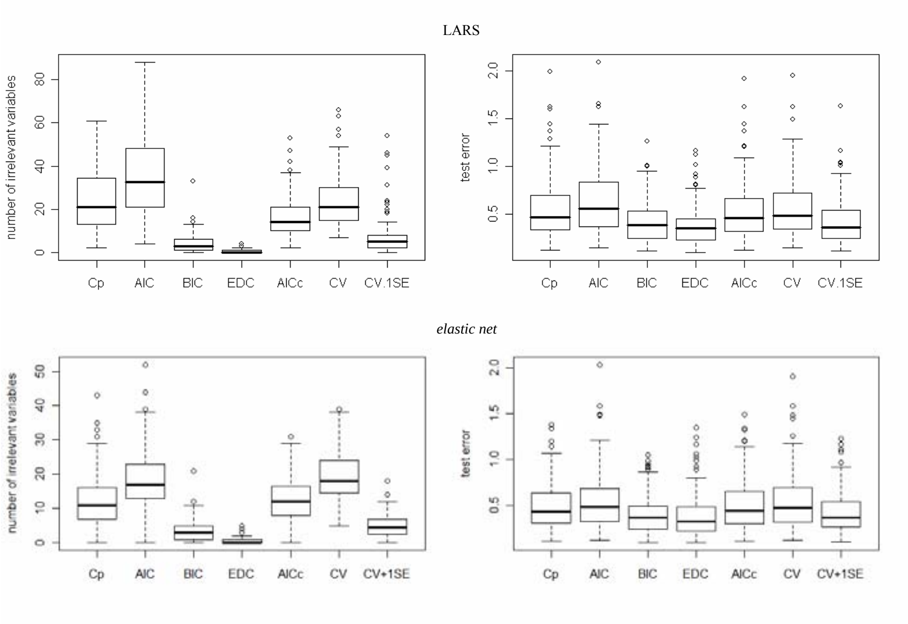

Figure 2. The comparison of seven model selection criteria in the case N=200 and p=100.
Source: own computations.

Experiment 3

This time we considered the case of only 50 observations in training sample and we added 30 noisy variables. The question is how model selection criteria deal with such settings with few degrees of freedom. The results are shown in Figure 3.

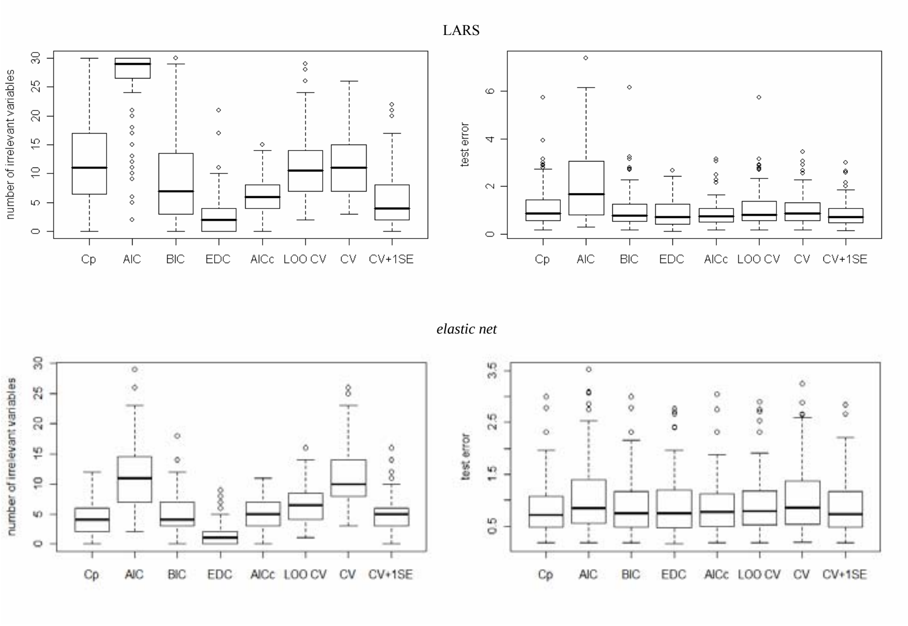

Figure 3. The comparison of eight model selection criteria in the case of few degrees of freedom (*N*=50 and *p*=40).

Source: own computations.

EDC is best in discarding noisy variables. The test errors seem to be comparable in the boxplot from right-down panel (for the majority of the criteria), but the null-hypothesis in Friedman test was rejected for both algorithms. The lowest average test error was yielded by C_p criterion (in *elastic net*) but the difference with error for EDC was not significant. The worst performance is seen in the case of the use of AIC.

V. CONCLUSIONS

Having conducted the experiments with various numbers of noisy variables we recommend the EDC criterion for model selection in LARS and *elastic net*. It clearly includes fewest noisy variables to the models. At the same time it yields prediction errors comparable to other criteria what was proved by post-hoc tests after Friedman test (the significance level was 0.05). Note that applying EDC criterion does not require the validation test, thus avoiding cross-validation procedure considerably reduces the computational cost in case of high dimension. The worst performance was observed for popular AIC criterion.

REFERENCES

Bai Z.D., Krishnaiah P.R., Zhao L.C. (1986), On the detection of the number of signals in the presence of white noise, *J. Multivariate Anal.* 20, p. 1–25.

Breiman L., Spector P. (1992), Submodel selection and evaluation in regression: the X-random case, *International Statistical Review* **60**: p. 291–319.

Burnham K. P., Anderson D.R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag.

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

Guyon I., Gunn S., Nikravesh M., Zadeh L. (2006), *Feature Extraction: Foundations and Applications.* Springer, New York.

Hastie T., Tibshirani R., Friedman J. (2009), *The Elements of Statistical Learning: Data Mining, Inferance, and Prediction.* 2nd edition, Springer, New York.

Hurvich C. M., Tsai C.-L. (1989), Regression and time series model selection in small samples, *Biometrika*, 76: p. 297–307.

Kundu D., Murali G. (1996), Model selection in linear regression, *Computational Statistics & Data Analysis* 22, p. 461–469.

Maddala G.S. (2008), *Ekonometria*, PWN, Warszawa.

Nemenyi P. B. (1963), *Distribution-free multiple comparisons*, PhD thesis, Princeton University.

Tibshirani R. (1996), Regression shrinkage and selection via the lasso, *J.Royal. Statist. Soc. B.*, 58: p. 267–288.

Wahba G. (1980), Spline bases, regularization, and generalized crossvalidation for solving approximation problems with large quantities of noisy data, *Proc. of the Inter. Conf. on Approximation theory in Honour of George Lorenz*, Academic Press, Austin, Texas, p. 905–912.

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

O WYBORZE POSTACI MODELU W WYBRANYCH METODACH REGULARYZOWANEJ REGRESJI LINIOWEJ

W ostatnich latach można zaobserwować dynamiczny rozwój różnych postaci regularyzacji w modelach liniowych. Wprowadzenie kary za duże wartości współczynników skutkuje zmniejszeniem wariancji (wartości współczynników są „przyciągane” do zera) oraz eliminacją niektórych zmiennych (niektóre współczynniki się zerują). Selekcja zmiennych za pomocą regularyzowanych modeli liniowych jest w problemach wielowymiarowych preferowana wobec popularnego podejścia polegającego na przeszukiwaniu przestrzeni cech i ocenie podzbiorów zmiennych za pomocą kryterium jakości modelu (*wrappers*). Przyczyną są mniejsze koszty obliczeń i mniejsza podatność na nadmierne dopasowanie. Jednakże wartości estymowanych współczynników (a więc także jakość modelu) zależą od parametrów regularyzacji. Zaimplementowane w tym celu w programie R kryteria jakości modelu nie gwarantują wyboru modelu optymalnego. Na podstawie przeprowadzonych symulacji w artykule proponuje się zastosowanie kryterium EDC.