Model selection based on combined penalties for biomarker identification

Eleni Vradi[1,2], Werner Brannath[2], Thomas Jaki[3], Richardus Vonk[4]

[1] Research and Early Development Statistics, Bayer AG, Berlin, Germany, [2] Institute for Statistics and Competence Center for Clinical Trials, Universität Bremen, Germany, [3] Medical and Pharmaceutical Statistics Research Unit, Lancaster University, UK, [4] Medical Writing and Statistics Oncology, Bayer AG, Berlin, Germany.

Contact: vradi.eleni@gmail.com

ABSTRACT

The growing role of targeted medicine has led to an increased focus on the development of actionable biomarkers. Current penalized selection methods that are used to identify biomarker panels for classification in high dimensional data, often result in highly complex panels that need careful pruning for practical use.

We propose a stepwise forward variable selection method which combines the L_0 with L_1 or L_2 norms. The penalized likelihood criterion that is used in the stepwise selection procedure results in more parsimonious models, keeping only the most relevant features. Simulation results and a real application show that our approach exhibits a comparable performance with common selection methods with respect to the prediction performance whilst minimizing the number of variables in the selected model resulting in a more parsimonious model as desired.

INTRODUCTION

In the biomarker research area, penalization techniques provide an attractive approach to the variable selection and are rather appealing due to their sparse representation and good predictive accuracy. In genomic research, an L_1 penalty [1] is routinely used due to its convexity and optimization simplicity. However, the result of the L_1 type regularization may not be sparse enough for a good interpretation. The development of methods to obtain sparser solutions is becoming essential part in the classification and feature selection area.

We proposed a stepwise forward approach for model selection in the framework of penalized regression using a penalty that combines the L_0 norm, which is based on the number of coefficients, with L_1 norm which is based on the size of coefficients or L_2 norm which take into account the grouping effect (i.e. keep in the model groups of variables that are correlated).

The aim is to find a model that includes as few relevant variables as possible with good predictive performance. This is an important consideration for classification and screening applications where the goal is to develop a test using as few features as possible.

METHODS

We consider the logistic regression model with logit link function ($logit(p) = X^T\beta$). The vector of *n* binary responses is $\mathbf{y} = (y_1, y_2, \dots, y_n)$ and $\mathbf{X}$ is an $n \times d$ matrix of predictors.

Under the regularization framework, the estimated coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_d) \in \mathbb{R}^d$ are obtained by minimizing the objective function

$$-logL + \lambda P(\boldsymbol{\beta})$$

where $P(\boldsymbol{\beta})$ is a regularization term and λ a tuning parameter.

The combined $L_0 + L_1$ penalty

Following Liu and Wu [2] the penalization term is defined as

$$CL_\alpha{}^\varepsilon(\beta) = (1-a)L_0{}^\varepsilon + aL_1$$

where $L_0{}^\varepsilon(\beta) = \begin{cases} 1, & |\beta| \geq \varepsilon \\ \frac{|\beta|}{\varepsilon}, & |\beta| < \varepsilon \end{cases}$ is a continuous approximation to L_0 and $0 \leq a \leq 1$

The estimated coefficients are obtained by minimizing the objective function

$$-logL + \lambda \sum_{j=1}^{d} CL_\alpha{}^\varepsilon(\beta_j)$$

However, the applicability of the proposed $CL_\alpha{}^\varepsilon$ penalty was restricted to moderate datasizes due to the nonconvexity of $CL_\alpha{}^\varepsilon$.

We consider a stepwise heuristic approach where at each step we obtain the estimated coefficients by:

$$\hat{\beta}_{L_1} = \underset{\beta}{argmin}\{-logL + \lambda\, a \sum_{j=1}^{d} L_1(\beta_j)\}$$

The selected model is based on the criterion that minimizes

$$-logL(\hat{\beta}_{L_1}) + \lambda a L_1(\hat{\beta}_{L_1}) + \lambda(1-a)L_0(\hat{\beta}_{L_1}) \quad \textbf{(1)}$$

Algorithm:

Given a set of d standardized predictors $X = X_1, \dots, X_d$, and a binary response $Y \in \{0,1\}$:

Step1: Consider all the univariate models

$$M_1: Y \sim \beta_1 X_1,\ M_2: Y \sim \beta_2 X_2,\ \dots,\ M_d: Y \sim \beta_d X_d$$

- Keep $M_j,\ j \in \{1, \dots, d\}$ that minimizes (1)

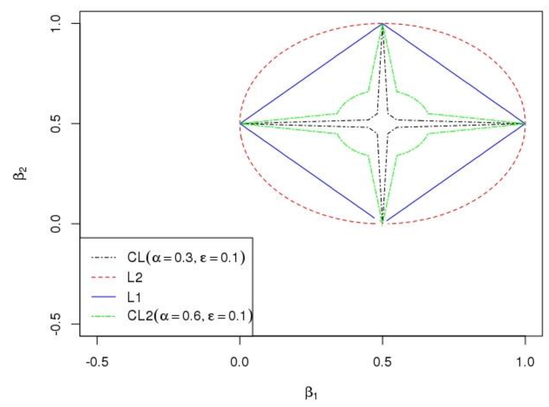
Step 2: With the model chosen in step 1, e.g. M_2, and in an additive way we consider all the $d-1$ models (M') by adding the remaining $d-1$ variables one at a time to the model M_2.

$$M'_1: Y \sim \beta_2 \mathbf{X_2} + \beta_1 X_1$$
$$M'_2: Y \sim \beta_2 \mathbf{X_2} + \beta_3 X_3$$
$$\vdots$$
$$M'_d: Y \sim \beta_2 \mathbf{X_2} + \beta_d X_d$$

- Keep the model that minimizes the function in (1)

Step 3: Continue until the value of the function (1) in the current step is bigger than the previous step.

Figure 1: Two-dimensional contour plot of $CL_{0.3}$, L_1, L_2, $CL2_{0.6}$ with $\varepsilon = 0.1$



RESULTS

We examine the performance of the proposed stepwise method with the above introduced combined penalties *(stepCL, stepCL2)* via simulations and a real data application. We include the global minimization *(CL, CL2)* for a comparison, and we also consider the results from Lasso and the adaptive Lasso. We compare the different methods

- in terms of model complexity
- the classification performance

All the functions that were used for the combined penalty approach can be found in the R-package "*stepPenal*", available on CRAN.

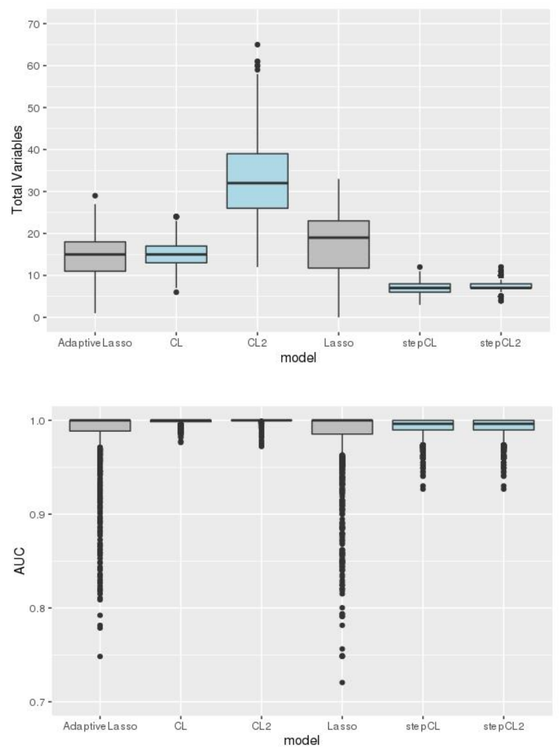
Tuning of parameters α, λ: 5-fold Cross Validation was used and the Area Under the ROC curve on the validation set is estimated over a grid of values of λ and α. The (λ ,α) that yields the 1sd from the maximum AUC is selected.

In this poster we will only present results from the real data application. Simulation results can be found in [2]

The real data come from a biomarker study with baseline measurements of *d=187* proteins from *n=53* patients. The objective is to extract potential candidate markers discriminating responders from non-responders.

In order to evaluate the performance of the models and in the absence of an external validation dataset we use bootstrapping. We applied all the methods on *B=1000* bootstrapped datasets of the protein data, by sampling with replacement. Figure 2 shows that the the classification performance of the stepwise methods *stepCL* and *stepCL2* is as good as the other variable selection methods (Lasso and adaptive Lasso), albeit including the least predictors.

Figure 2: Boxplots of the total Variables selected by the models (top panel) and the AUC of the ROC curves (bottom panel) over the B=1000 bootstrapped data of the protein data



CONCLUSION

We have proposed a stepwise forward approach for model selection in the framework of penalized regression using a penalty that combines the L_0 with L_1 or L_2 norms. The aim is to find a model that includes as few relevant variables as possible on the one hand, and have good predictive performance on the other hand. The combined penalization term $CL_\alpha{}^\varepsilon(\beta)$ that was introduced by Liu and Wu [3] was limited to moderate datasets due to limitations of the optimization algorithm. Considering the heuristic stepwise forward approach, we can apply the penalization $CL_\alpha(\beta)$ and $CL2_\alpha(\beta)$ to high-dimensional data by using the BFGS algorithm in a stepwise approach which is found to work well in practice for nonconvex and non-smooth functions [4]. As future work we also consider to apply our method to regression problems for variable selection with a continuous response as well as time-to-event data.

REFERENCES

1. Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *J.R. Statist. Soc. B.*
2. Vradi, E., Brannath, W., Jaki, T. and Vonk, R., 2018. Model selection based on combined penalties for biomarker identification. *Journal of biopharmaceutical statistics*, *28*(4), pp.735-749.
3. Liu,Y. and Wu,Y., 2007.Variable selection via a combination of the L0 and L1 penalties. *Journal of Computational and Graphical Statistics*, *16*(4).
4. Curtis,F.E. and Que,X., 2015. A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees. *Mathematical Programming Computation*, *7*(4).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 633567 and is part of the IDEAS European training network *(http://www.ideas-itn.eul)*.