Model selection based on combined penalties for biomarker identification

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, however, often result in highly complex panels that need careful pruning for practical use. In the framework of regularization methods a penalty that is a weighted sum of the L_1 and L_0 norm has been proposed to account for the complexity of the resulting model. In practice, the limitation of this penalty is that the objective function is non-convex, non-smooth, the optimization is computationally intensive and the application to high-dimensional settings is challenging. In this paper 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.

Keywords: biomarker panels, combined penalties, model selection, penalized regression, regularization, sparsity, stepwise variable selection, treatment responder.

1. Introduction

The high costs and long duration of clinical development, paired with high levels of attrition, require the quantification of the risk when moving from early to late stage clinical development, and biomarkers may play an important role in this quantification. However, only rarely the number of variables (biomarkers) in the resulting panel plays an active role in selection procedures. Variable selection is an important aspect in the determination of such panels in the

framework of high-dimensional statistical modeling. In practice, a large number of candidate predictors are available for modeling. Keeping only the relevant variables in the model makes interpretation easier and may increase the predictability of the resulting model.

Particularly in the framework of regularization methods, various penalty functions are used to perform variable selection. Frank and Friedman (1993) proposed the bridge regression by introducing the penalty of the form $L_q = \sum_{j=1}^{d} |\beta_j|^q, q > 0$, for the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d) \in \mathbb{R}^d$. When $q \leq 1$ the penalty performs variable selection. The case where $q = 1$ is the L_1 penalty and corresponds to the Least Absolute Shrinkage and Selection Operator (Lasso) (Tibshirani, 1995). It performs continuous shrinkage and variable selection at the same time, whereas for $q = 2$ we get the ridge estimator (Hoerl and Kennard, 1970) that shrinks coefficients towards zero but it does not perform variable selection. The limit of the L_q as $q \to 0$ gives the L_0 penalty, which penalizes the number of non-zero coefficients and thus is appealing for model selection, if sparse models are of advantage. However, due to its non-convexity and discontinuity at the origin, the corresponding optimization problem becomes difficult to implement in high dimensions. In addition, the solution using L_0 may be unstable because it may not be identifiable.

In genomic research, an L_1 penalty 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 than through L_1 penalization methods is becoming essential part in the classification and feature selection area. A variable selection method that combines the L_1 and L_0 penalties was proposed by Liu and Wu (Liu and Wu, 2007). They used a mixed integer programming algorithm for optimization of the objective function. The results showed that their method achieved sparser solutions than Lasso and more stable solutions that the L_0 regularization. However the application was limited to

moderate data sizes, due to computational inefficiency for large-scale problems. Other combinations of L_q penalties have been proposed so far (Zou and Hastie, 2005) and recently (Huang et al., 2016) with each of these methods using a different optimization algorithm to approach the solution.

In this article, we propose a method for variable selection that penalizes the likelihood function with a linear combination of L_0 with L_1 or L_2 penalties (*CL, CL2*) in a stepwise forward variable selection procedure. The aim is to obtain a model that is sparser than the model with the L_1 penalty alone and at the same time achieve a good predictive performance. Moreover, a strong motivation for the proposed stepwise forward variable selection method is that state-of-the-art global optimization algorithms for non-smooth and nonconvex functions do not provide satisfactory results. In section 2, we define the *CL* and *CL2* penalties and present the algorithm for solving the penalized logistic regression problem with these combined penalties. In section 3, we use simulated data to evaluate the performance of our method and we compare it to Lasso and adaptive Lasso both in terms of correct variable selection (true covariates with $\beta_j \neq 0$) as well as predictive performance. Finally, we show an application of our method for classification and variable selection on a real dataset with protein measurements to identify the least number of predictors that can best classify responders and non-responders to a treatment.

2. Methods

2.1 Regularization

Suppose we have data (**X, y**), where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is the vector of responses and **X** is an $n \times d$ matrix of predictors. We will assume that the observations are independent and the predictors standardized. With linear predictor $\eta = \boldsymbol{X^T\beta}$ and link function g the generalized linear model is expressed as

$$g(E(\boldsymbol{y}|\boldsymbol{X})) = \eta \qquad (2.1)$$

Under the regularization framework, the estimated coefficients $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_d\right) \in \mathbb{R}^d$ are obtained by minimizing the objective function $-logL + \lambda P(\boldsymbol{\beta})$, and are given by:

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{argmin}\{-logL + \lambda P(\boldsymbol{\beta})\}$$

where $P(\beta)$ is a regularization term. The parameter λ>0 is a tuning parameter and $-logL$ is the negative log-likelihood. One of the most popular and commonly used regularization method is the L_1 regularization (Lasso), where $P(\beta) = \sum_{j=1}^{d}|\beta_j|$. Setting $\lambda = 0$ reverses the Lasso to Maximum likelihood estimation. On the other hand, a very large λ will completely shrink β to zero thus leading to the empty or null model. In general, moderate values of λ will cause shrinkage of the solutions towards zero, and some coefficients may be exactly zero.

Other types of L_1 regularization include the adaptive Lasso, where adaptive weights are used for penalizing different coefficients in the L_1 penalty and was shown to have the oracle property (Zou, 2006). A variable selection and estimation procedure is said to have the oracle property i) if it selects the true model with probability tending to 1 and ii) if the estimated penalized coefficients are asymptotically normal, with the same asymptotic empirical variance as the estimator based on the true model.

However, the L_1 type regularization is consistent only under rather restrictive assumptions (Zhao and Yu, 2006) and the coefficient estimates are severely biased due to shrinkage (Meinshausen and Yu, 2009); (Fan and Li, 2001). Although the L_0 norm, where $P(\beta) = \sum_{j=1}^{d} 1_{\beta_j \neq 0}$ and $1_{\beta_j \neq 0}$ is the indicator function of whether $\beta_j \neq 0$, tend to yield the sparsest solutions, its implementation in high dimensional data becomes an NP hard optimization problem and is not computationally feasible. Classical information criteria like AIC (Akaike, 1974) or BIC (Schwarz, 1978) lie in the general class of the regularization $\lambda P(\beta) =$

$\lambda \sum_{j=1}^{d} 1_{\beta_j \neq 0}$ for suitable choices of λ. In order to gain a more concise and sparse solution and whilst keeping a high predictive accuracy of the classification model, we propose a regularization term that combines the $L_0$with L_1 or L_2 norms (Liu and Wu, 2007). Figure 1 plots the penalty functions L_1 and L_2 in the bottom panel and the L_0 penalty in the top panel. Unlike L_2 , the penalty terms L_1 and L_0 are singular at the origin and thus perform variable selection (Fan and Li, 2001).

Figure 1.

2.2 The combined $L_0 + L_1$ penalty

Following Liu and Wu (Liu and Wu, 2007) the penalization term is defined as $CL_\alpha{}^\varepsilon(\beta) = (1-a)L_0{}^\varepsilon + aL_1$, where $0 \leq a \leq 1$ is a weighting parameter between $L_0{}^\varepsilon$ and L_1 penalties, with $L_0{}^\varepsilon$ given by:

$$L_0{}^\varepsilon(\beta) = \begin{cases} 1, & |\beta| \geq \varepsilon \\ \frac{|\beta|}{\varepsilon}, & |\beta| < \varepsilon \end{cases} \tag{2.2}$$

Clearly $CL_1{}^\varepsilon = L_1$ $(a = 1)$ and $CL_0{}^\varepsilon = L_0{}^\varepsilon$ $(a = 0)$ are special cases of $CL_\alpha{}^\varepsilon$. Discontinuity at the origin of L_0 makes the optimization difficult and therefore we consider the continuous approximation to L_0 defined by (2.2). The limit of $L_0{}^\varepsilon(\beta)$ when $\varepsilon \to 0$ is $L_0(\beta)$ itself. When $\varepsilon > 0$ is small, $L_0{}^\varepsilon(\beta)$ is a good approximation to $L_0(\beta)$ (Figure 1 top right). The estimated coefficients are obtained by minimizing the objective function

$$-logL + \lambda \sum_{j=1}^{d} CL_\alpha{}^\varepsilon(\beta_j) \tag{2.3}$$

and are given by

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

2.3 The combined $L_0 + L_2$ penalty

We now consider another combination, the L_0 norm with L_2 . The motivation for combining the L_0 norm with L_2 , is to consider a penalty that will join the nice properties of the L_2 and those of the L_0 norm, which is to perform variable selection (L_0) and keep in the model groups of variables that are correlated (L_2). In theory, a strictly convex function provides a sufficient condition for such grouping of variables and the L_2 penalty guarantees strict convexity. The grouping effect refers to the simultaneous inclusion (or exclusion) of correlated predictors in the model.

The penalization term is now defined $CL2_\alpha{}^\varepsilon(\beta) = (1-a)\, L_0{}^\varepsilon + a\, L_2$, where $0 \le a \le 1$. The $L_0{}^\varepsilon$ term introduced above is for variable selection and the L_2 penalty shrinks the coefficients towards zero with no contribution to variable selection. Figure 2 gives a graphical representation of the regularization terms $CL_{0.3}{}^{0.1}$, L_1 , L_2, $CL2_{0.5}{}^{0.1}$.

Figure 2.

2.4 The stepwise forward procedure

In their paper Liu and Wu (2007) proposed a global algorithm to solve the corresponding difficult nonconvex problem (Mixed integer linear programming). However, the applicability was restricted to moderate datasizes. As mentioned by Frommlet F. and Nuel G. (Frommlet and Nuel, 2016), when the number of predictors grows large (*d>20*) it is not possible to apply algorithms which guarantee to find the optimal solution (Furnival and Wilson, 2000) and instead heuristic search strategies like stepwise procedures may be considered. By using heuristic techniques, we can approximate the solution of the non-smooth, non-convex and NP-hard optimization problems like the one in equation (2.3), where exact algorithms are not applicable for such minimization problems.

The optimization of the objective function (2.3) is rather challenging since ${CL_\alpha}^{\varepsilon}(\beta)$ and ${CL2_\alpha}^{\varepsilon}(\beta)$ are non-convex and non-differentiable at some points of the parameters' space. We apply the BFGS, Broyden (Broyden, 1970)- Fletcher (Fletcher, 1970)- Goldfarb (Goldfarb, 1970)- Shanno (Shanno, 1970) (BFGS) variable metric (quasi Newton) method, which is shown to work well for the optimization of non-smooth and non-convex functions (Lewis and Overton, 2009); (Lewis and Overton, 2013); (Curtis and Que, 2015). The BFGS method uses an approximation of the Hessian matrix to find the stationary points of the function to be minimized. Its ability to capture the curvature information of the considered function makes the method so effective.

We propose to use a stepwise forward variable selection using the previously introduced penalized likelihood criterion for feature selection that can be used effectively in high dimensional data. In this stepwise forward selection framework, at each step we optimize the objective function $-logL + \lambda\, a\, L_1$ using the BFGS algorithm and obtain

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

The selected model is based on the criterion that minimizes the value of

$$-logL(\hat{\beta}_{L_1}) + \lambda\, a\, L_1(\hat{\beta}_{L_1}) + \lambda\,(1-a)\, L_0(\hat{\beta}_{L_1}) \qquad (2.4)$$

The suggested algorithm is described as follows:

Step 1:

- Given a set of d standardized predictors $X_1, X_2, \dots, X_d$ and a response $y_i \in \{0,1\},\ i = 1, \dots n$ we consider all possible univariate models $(M_1, M_2, \dots, M_d)$

$$M_1\colon\ Y \sim \beta_0 + \beta_1 X_1,\quad M_2\colon Y \sim \beta_0 + \beta_2 X_2,\quad M_3\colon\ Y \sim \beta_0 + \beta_3 X_3, \dots,\ M_d\colon\ Y \sim \beta_0 + \beta_d X_d$$

- Estimate ${\hat{\beta}_{L_1}}^{M_1}, \dots, {\hat{\beta}_{L_1}}^{M_d}$ and keep $M_j, j \in \{1, \dots, d\}$ that gives the smallest value of the function (2.4), e.g. keep variable X_2

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.

- $M'_1 : \mathrm{Y} \sim \beta_0 + \beta_2 \mathbf{X_2} + \beta_1 X_1$

  $M'_2 : \mathrm{Y} \sim \beta_0 + \beta_2 \mathbf{X_2} + \beta_3 X_3$

  $\vdots$

  $M'_d : \mathrm{Y} \sim \beta_0 + \beta_2 \mathbf{X_2} + \beta_d X_d$

- Keep the model that minimizes the function in (2.4),

Step 3:

- Continue adding single variables until the value of the function (2.4) in the current step is bigger than its value in the previous step.

The advantage of using the function (2.4) instead of (2.3) in the optimization is that we no longer need to consider the continuous approximations to the discontinuous L_0 function and therefore we eliminate the number of parameters by the continuity parameter ε**.** The reason why we can do so is that within each step the L_0-penality term remains constant (since the dimension of the model is fixed) and hence play no role in the determination of the regression coefficients. The L_0-penality term does only play a role for the stopping criterium.

2.5 Sparse logistic regression with combined penalties

As a particular example we consider the binary linear regression model (2.1), where $y \in \{0,1\}$, is a vector of n observed binary outcomes, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_d) \in \mathbb{R}^d$ is the vector of coefficients. The link function is the logit function $logit(p) = \log\left(\frac{p}{1-p}\right)$, where p is the conditional event probability and is given by

$$p = P(\boldsymbol{y} = 1|\boldsymbol{X}) = \frac{e^{\eta}}{1+e^{\eta}}$$

(2.5)

The coefficient estimates are obtained by minimizing (2.3) with the log-Likelihood

$$logL = L(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} y_i \log(p_i) + (1 - y_i)\log(1 - p_i)$$

3. Results

In this section we examine via simulations the performance of logistic regression when models are selected and estimated with the above introduced combined penalties (*CL, CL2)* by either stepwise forward selection as introduced in Section 2 *(stepCL and stepCL2)* or by global minimization (*CL, CL2)*. In the stepwise model selection scheme, we also examine the performance of the stepwise adaptive L_1 model with the $\lambda\ (1 - a)\ L_0$ selection criterion *(stepAdaCL)*. For that model the objective function to minimize is $-logL + \lambda\ a\ \sum_{j=1}^{d} w_j L_1(\beta_j)$, where $w_j = \frac{1}{|\beta_j^*|^{\gamma}}$ are the adaptive weights and $|\beta_j^*|$ is the ridge regression estimator. The estimation is done with the stepwise algorithm described in Section 2.4.

Although the proposed method is a stepwise variable selection procedure, we did not consider the comparison with other stepwise methods like the stepwise BIC or AIC, as they tend to perform poorly when the dimension is large relative to the sample size and are usually too liberal, that is, they tend to select a model with many spurious covariates (Chen and Chen, 2008). As mentioned by Zhang and Shen (Zhang and Shen, 2010) these criteria may be inadequate due to their nonadaptivity to the model space and infeasibility of exhaustive search.

We include the global minimization in spite of the disadvantages mentioned in Section 2 for a comparison. We also consider the results from Lasso (L_1 penalty) and the adaptive Lasso. We compare the different methods in terms of the fraction of correctly selected variables and the

prediction classification accuracy. The real data come from a biomarker study concerned with protein measurements with the objective to select biomarkers that potentially discriminate between responders and non-responders.

3.1 Simulation Study

We simulate data for varying number of predictors. We consider two settings, one high dimensional data where the number of predictors (*d*) exceed the number of samples (*n*), and a setting where the sample size is smaller than the dimensionality of the data. We assume multivariate normal predictors $X_1, \dots, X_d$ with pairwise correlation ρ (compound symmetry). Let ρ denote the correlation between variables $X_m,\ X_l$ where $m, l\ \in \{1, \dots, \mathrm{d}\}$, $m \neq l$.

The true model that was used to generate the outcome has *k* informative covariates $X_k,\ k\ \in \mathbb{Z}$, $1 < k < d$. We consider a classification problem with $\boldsymbol{y}$ a binary response and standard normally distributed predictors $\boldsymbol{X} \sim MVN(0, \Sigma)$, where Σ is the covariance matrix. We consider the logistic model with logit link function, $logit(p) = X^T\beta$, as described above with p the probability of $y=1$ given X as defined in (2.5). In other words, each component of the response vector *y* is viewed as a realization of a Bernoulli random variable with probability of success p.

Four scenarios will be presented here, each with *n=100* samples.

1. Scenario 1: $d<n$, correlation $\rho = 0.5$

We consider *d=50* covariates, with *k=3* informative predictors and coefficient vector $\beta = (3, 1.5, 2, \underbrace{0, \dots, 0}_{47})$. The correlation between the 3 informative predictors is $\rho = corr(X_m, X_l) = 0.5$, $m \neq l$ and $m, l = 1,2,3$.

2. Scenario 2: High dimensional setting $d>n$, correlation $\rho = 0.5$

We consider *d=150* covariates, with *k=15* informative predictors with $\beta =$ $(\underbrace{3,\dots}_{3},\underbrace{-3.5,\dots}_{3},\underbrace{1.5,\dots}_{3},\underbrace{5,\dots}_{4},\underbrace{-2,\dots}_{2},\underbrace{0,\dots,0}_{135})$. The correlation ρ between the 15 informative predictors is $\mathrm{corr}(X_m, X_l) = 0.5,\ m \neq l$ and $m, l = 1,\dots,15$.

3. Scenario 3: High dimensional setting *d>n,* correlation $\rho = 0.7$

The dataset consists of *n=100* samples and *d=200* covariates, with *k=15* informative predictors with $\beta = (\underbrace{2,\dots}_{4},\underbrace{-3,\dots}_{3},\underbrace{1.5,\dots}_{4},\underbrace{-2,\dots}_{4},\underbrace{0,\dots,0}_{185})$. The correlation ρ between the 15 informative predictors is $\mathrm{corr}(X_m, X_l) = 0.7,\ m \neq l$ and $m, l = 1,\dots,15$

4. Scenario 4: High dimensional setting *d>n,* block correlation

We consider *d=200* covariates, with *k=16* informative predictors with $\beta =$ $(1,\underbrace{4,\dots}_{4},\underbrace{-3,\dots}_{3},\underbrace{1.5,\dots}_{4},\underbrace{-2,\dots}_{4},\underbrace{0,\dots,0}_{184})$. In this scenario there are two groups (blocks) of correlated predictors and one single independent feature. The coefficients of *d-k*=184 variables were set to zero, $\beta_r = 0,\ r = 184,\dots,200$. The correlation between predictors in block 1 is $\mathrm{corr}(X_m, X_l) = 0.4,\ m \neq l$ and $m, l = 1,\dots,7$ and the correlation among predictors in block 2 is $\mathrm{corr}(X_m, X_l) = 0.7,\ m \neq l$ and $m, l = 8,\dots,16$.

3.2 Tuning of parameters

All analyses were done in R version 3.2.3 (R Core Team, 2015). For the Lasso and Adaptive Lasso the *glmnet* library was used and all the functions that were used for the combined penalty approach can be found in the R-package "*stepPenal*", available on the CRAN. For the adaptive lasso weights were estimated by ridge regression and then used for a weighted L_1 penality in estimation of β. The optimal regularization parameters for the methods *stepCL, stepAdaCL, CL, CL2, stepCL2* were tuned by 10-fold cross-validation on the two dimensional surface (a, λ) using a grid of values. The choice of the optimal parameters was done in the following way. For

each configuration of (a, λ) in the grid, the AUC of the ROC curves on the validation set was computed in each of the 10 validation sets. The average of the 10 AUCs was reported together with its standard deviation.

Selection of (a, λ) was based on the interval $A = [maxAUC - sdAUC, maxAUC)$ where maxAUC is the maximum average AUC and sdAUC is the standard deviation of the AUCs corresponding to the (a, λ) with maximum average AUC. The (a, λ) that corresponds to the median of the AUCs in the interval A was chosen for the final model fitting. In case that more than one configurations yields the median of the AUCs, we select the configuration with the largest λ and smallest a, to obtain the sparsest model. The use of the interval A acknowledges the sample variability and the fact that we are aiming for a compromise between good classification performance and complexity of the model. In other words, a small decrease in the AUC of the ROC curve is acceptable in return to a less complex model. The Lasso and adaptive lasso were also tuned by 10-fold cross-validation on the one dimensional space (λ), using the default settings in R in the function *cv.glmnet* and the measure type *"auc"*.

3.3 Simulation Results

The different classifiers were built by the estimated tuning parameters on the training set. Then, the obtained classifiers were applied to the testing set for classification and prediction. For the testing set, we simulated data from the same distribution as the training set for *n=1000* samples. We simulated 1000 datasets on which we applied all methods. For each method we computed the mean classification performance of the models on the testing sets measured by the AUC of the ROC curve (test AUC). This is a measure for the discrimination ability of the model to correctly distinguish the two classes of the response. The complexity of the resulting model was measured by the ratio of correctly selected variables (true covariates with $\beta_j \neq 0$) to the total variables selected by the model. We will call this ratio *RCV* for the rest of the paper.

This ratio takes values between zero and one. When the model selects none of the informative variables it is zero and it becomes one, when the selected model includes only the *k* informative covariates. The closer the *RCV* is to one, the sparser the model is and selects the true variables. The results in Table 1 summarize the performance of the different methods in terms of model complexity. An ideal model selection method would only select the *k* true features and set the coefficients of the other *d-k* variables equal to zero.

Table 1:

In most of the scenarios, the *stepCL* and *stepCL2* methods yield a higher *RCV* than the other methods and on average the *stepCL* and *stepCL2* models are sparser than the other methods. In scenario 2 and 3 the adaptive Lasso yields the higher *RCV*, but the models are not as sparse as the stepwise methods. Although the stepwise methods (*stepCL, stepCL2, stepAdaCL*) result in including the least variables in the model, its discriminative ability in terms of AUC, as shown in Table 2, is comparable with the other methods that tend to select a larger model with more variables.

The *stepCL2* method also has remarkable performance both in terms of sparsity and predictive discrimination. Considering the trade-off between model complexity and performance, the proposed stepwise combined penalty approach achieves a good balance between parsimony, including less variables and maintaining a high predictive accuracy that is comparable with state-of-the-art methods. We should mention that in scenarios 2,3 and 4 none of the methods select all of the *k* informative variables, however, for the stepwise method the AUC of the ROC curve on the testing set is greater than 90%, indicating a good discrimination accuracy by including the least variables in the model. In all scenarios, we found that the *stepCL2* method has comparable performance to *stepCL* and is superior to adaptive Lasso and Lasso.

In Table 2 we present results regarding the predictive classification accuracy of the methods by the AUC of the ROC curves. We report the Brier score (Brier, 1950) as a measure of the accuracy of predictions, defined as

$$Brier = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2$$

It is given by the squared distance between the patients observed status y_i and the predicted probability $\hat{p}_i$. The decision space for the Brier score is the interval [0,1] and generally the lowest the Brier score , the better the classification rule. If the predicted probability is 0.5 for each individual, the Brier score of 0.25 would indicate that the classification rule is a random one.

Table 2:

Empirical results from our simulations show that even for no high-dimensional settings where $n>d$, the stepwise method gives the sparsest solutions whilst maintaining classification performance measures as good as Lasso and adaptive Lasso. The *CL2* method tends to select big models, due to the L_2 norm which shrinks coefficients towards zero without variable selection. Thus when a is close to 1, the model will behave similar to ridge regression and the resulting model will be complex in terms of the number of predictors. On the other hand, when a is closer to 0, the *CL2* and *stepCL2* penalties will borrow more of the characteristics of the L_0 norm and will result in sparser models.

In our simulations we also considered the case where there is no correlation among predictors (results not shown in the table as we don't consider it a realistic scenario). We repeated scenario 2 with the only alteration of setting $\rho = 0$. Results were in the same direction as in scenario 2 shown in Table 1 and Table 2. That is, the stepwise methods perform better than all the other

methods in terms of model complexity resulting in the sparsest models with a high classification performance.

Furthermore, we examined the situation where there are no predictors in the data associated with the outcome. In that case that the true model is the null model, none of the methods identified the true model. Again, running through again the second scenario with *d=150* predictors with none being informative for the outcome, the *stepCL* method selected a median of 5 variables whereas the other methods selected between 14 (*AdaLasso, CL*) and 19 (*CL2*). We observed the same pattern in the results for repeating the first scenario with *d=50* uninformative predictors, where none of the methods selected the true model but the stepwise methods produced the sparsest solutions.

3.4 Application- real data analysis

To illustrate the applicability of the proposed method, we applied the stepwise method on a real example involving protein measurements. The dataset contained *n=53* patients with baseline measurements of *d=187* proteins. To maintain confidentiality, the names of the proteins are not revealed. For the presentation of the results and keeping the study anonymized we renamed the proteins to $X_1, X_2, \ldots, X_{187}$. The objective is to extract potential candidate markers discriminating responders from non-responders based on patients' protein levels. We apply our proposed stepwise combined penalty approach with the aim to select a small set of proteins that can sufficiently predict response to the treatment. We compare our approach with the commonly used Lasso and adaptive Lasso, but also with the global optimization penalized methods *CL* and *CL2*.

The regularization parameters were not tuned by cross-validation, due to the relatively small sample size (*n=53*). The tuning was done using the bootstrap method in the following way; for a

grid of values of (a, λ), we trained the models on *B=100* bootstrapped datasets (drawing samples with replacement from the original data) and evaluate their classification performance (in terms of AUC) on the original data. For each combination of the tuning parameters (a, λ), the models were trained on B bootstrapped sets and validated on the testing set (original data) and the average AUC (over the B bootstrapped samples) was reported together with its standard deviation. The configuration of (a, λ) that corresponds to the median of the AUC in interval *A*, as described above in the section 3.2 'Tuning of parameters', was chosen.

The results show that the stepwise methods yield the sparsest models by selecting 8 variables (*stepCL)* and 9 (*stepCL2*) accordingly, whereas the other methods select between 22 (*CL2*) and 26 (*Lasso*). It is noticeable that the classification performance of the stepwise method is as good as the other variable selection methods, albeit including the least predictors. 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 another *B=1000* bootstrapped datasets of the protein data, by sampling with replacement, and the frequencies of the top 10 selected variables by all methods are reported in Figure 3. This results in 16 unique proteins.

This figure shows that the proteins that were frequently selected by the *stepCL* and *stepCL2* methods are also frequently selected by the Lasso and adaptive Lasso. Note that the stepwise methods have lower frequencies of the selected variables, because selection of larger models will automatically increase the number of selection for individual variables.

Figure 3:

Figure 4 shows boxplots of the total number of variables included in the model over the bootstrap evaluations. The stepwise method yields consistent model selection by selecting a

median of 8 variables for *stepCL* and *stepCL2*, whereas the Lasso and Adaptive Lasso have a big variability on the complexity of the model selected. The AUC of the ROC curves that is used as a measure of classification performance of the methods on the bootstrapped datasets and their distribution is shown in Figure 5. The stepwise methods tend to always select the most sparse models more systematically, while maintaining a very good classification performance.

Figure 4:

Figure 5:

4. Conclusion

In this paper we have 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. The aim of the proposed method is to find a model that includes as less and relevant variables on 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 (2007) 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 which is found to work well in practice for nonconvex and non-smooth functions (Lewis and Overton, 2009); (Lewis and Overton, 2013); (Curtis and Que, 2015). As a result, the practical implementation of the stepwise penalization method is simpler and more efficient.

We found that for the stepwise method the computational time was shorter than the global optimization. However, the tuning of the regularization parameters (a, λ) can be computationally intensive. This is an important aspect of penalization methods and can be

453 further explored and improved in future work. Simulation results and a real data application
454 show that the proposed method yields sparser models, while maintaining a good classification
455 performance. This is an important consideration for classification and screening applications
456 where the goal is to develop a test using as less features as possible to control the cost. Overall,
457 we found that our method provides a sparser model whilst maintaining similar prediction
458 properties with the other methods. We hope that this paper could be a first step to learn more
459 about the theoretical properties of this method, which seems to be worth of further investigation.
460
461 Furthermore, it would be of great interest to extend the forward stepwise method to the stepwise
462 bidirectional approach, considering at each step of the algorithm which variables can be
463 included and excluded (forward and backwards variable selection) in the model. As future work
464 we also consider to apply our method to regression problems for variable selection with a
465 continuous response as well as time-to-event data.
466
467 **Declaration of Conflicting Interests**
468
469 The author(s) declared no potential conflicts of interest with respect to the research, authorship,
470 and/or publication of this article.
471
472 **Funding**
473
474 This project has received funding from the European Union's Horizon 2020 research and
475 innovation programme under the Marie Sklodowska-Curie grant agreement No 633567 and is
476 part of the IDEAS European training network (***http://www.ideas-itn.eu/***). This report is in part
477 independent research arising from Prof Jaki's Senior Research Fellowship (NIHR-SRF-2015-
478 08-001) supported by the National Institute for Health Research. The views expressed in this
479 publication are those of the authors and not necessarily those of the NHS, the National Institute
480 for Health Research or the Department of Health.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* p. 19(6): 716–723.

Brier, G., 1950. Verification of forecasts expressed in terms of probability. *Mon Wea Rev.,* pp. 78;1-3.

Broyden, C. G., 1970. The convergence of a class of double-rank minimization algorithms.. *IMA Journal of Applied Mathematics 6(1),* pp. 76-90.

Chen, J. and Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, pp.759-771.

Curtis, F.E. and Que, X., 2015. A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees. *Mathematical Programming Computation*, *7*(4), pp.399-428.

Fan, J. and Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, *96*(456), pp.1348-1360.

Fletcher, R., 1970. A new approach to variable metric algorithms. *Compt.J 13(3),* pp. 317-322.

Frank, L.E. and Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), pp.109-135.

Frommlet, Florian, and Gregory Nuel. "An Adaptive Ridge Procedure for L 0 Regularization." *PloS one* 11, no. 2 (2016): e0148620.

Furnival, George M., and Robert W. Wilson. "Regressions by leaps and bounds." *Technometrics* 16, no. 4 (1974): 499-511.

Goldfarb, D., 1970. A family of variable-metric methods derived by variational means. *Mathematics of computation 24(109),* pp. 23-26.

Hoerl, A.E. and Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), pp.55-67.

Huang, H.H., Liu, X.Y. and Liang, Y., 2016. Feature Selection and Cancer Classification via Sparse Logistic Regression with the Hybrid L 1/2+ 2 Regularization. *PloS one*, *11*(5), p.e0149675.

Lewis, Adrian S., and Michael L. Overton. "Nonsmooth optimization via BFGS." *Submitted to SIAM J. Optimiz* (2009): 1-35.

Lewis, A.S. and Overton, M.L., 2013. Nonsmooth optimization via quasi-Newton methods. *Mathematical Programming*, pp.1-29.

Liu, Y. and Wu, Y., 2007. Variable selection via a combination of the L 0 and L 1 penalties. *Journal of Computational and Graphical Statistics*, *16*(4), pp.782-798.

Meinshausen, N. and Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, pp.246-270.

R Core Team, 2015. *R: A Language and Environment for Statistical Computing,* Vienna, Austria: R Foundation for Statistical Computing.

Schwarz, G. E., 1978. Estimating the dimension of a model. *Annals of Statistics,* p. 6 (2): 461–464.

Shanno, D. F., 1970. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation, 24(111),* pp. 647-656.

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. *J.R. Statist. Soc. B,* pp. 267-288.

Zhang, Y. and Shen, X., 2010. Model selection procedure for high-dimensional data. *Statistical analysis and data mining*, *3*(5), pp.350-358

Zhao, P. and Yu, B., 2006. On model selection consistency of Lasso. *Journal of Machine learning research*, *7*(Nov), pp.2541-2563.

Zou, H., 2006. The adaptive lasso and its oracle properties. *J Am Stat Assoc. Taylor & Francis,* pp. 101:1418-1429.

Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), pp.301-320.