Analysis of Feature-Selection for LASSO Regression Models

Johannes Giersdorf          Miro Conzelmann

July 31, 2017

Abstract

To extract features from large data sets is a major issue in data science. Many statistical regression models exist that address this problem with many different approaches. In this project models belonging to the LASSO family are analyzed with respect to their ability and performance on feature selection in the domain of high dimensional data. LASSO models are regularized regression models aiming to create sparse models. The Aim is to develop methods to describe and compare the Feature-Selection behavior of these models with an empirical approach.

Keywords: LASSO – Adaptive LASSO – Group LASSO – Feature-Selection

1 Introduction

Regularized statistical modelling is a very popular approach to increase the performance of regression models that are aimed at local error minimization e.g. least square minimization. The idea of a penalized model (ridge regression) was introduced by Arthur et al. 1970 [4] to create more complex and flexible models have shown an increased prediction performance and were applied in various fields of regression problems. In turn ridge regression has been adapted and extended by Tibshirani, 1996 [10] in the way that the penalty is done in the L1 Norm. This approach is described by the least absolute shrinkage and selection operator (LASSO) and has opened a new branch of models. In this paper we will focus on the "Group LASSO" approach by Yuan and Lin, 2006 [11] which takes into account that groups of variables can be treated as single factors and the "adaptive LASSO" approach by Hui Zou 2006 [13] who introduces a weighting for each variable to improve selection consistency. The idea is to compare these models with regard to their consistency in variable selection within 1000 independent 10-fold random Crossvalidations (CV) as basis for the model selection. Each run consists of three CVs. At first a CV with standard LASSO is performed and the fold-selection sequence was stored. Then the other models followed but this time the CV used the formally saved folds of the standard LASSO CV. This procedure yields an optimal model for each LASSO-model and each run. The underlying data is a real-world data set of EEG measurements based on a multi-centre study on smoking-related behavior [6] and contains a binomial factor smoker(smk)/nonsmoker(nsmk), which was not analyzed regarding feature selection before. So the data set is treated like a random set, but with no information about the generation of the (random) numbers. This leaves us with the task to find an appropriate method to describe the behaviour of the LASSO-models applied to an unknown real-world application. It is noted that we are therefore not analyzing the sign consistency described by Peng Zhao and Bin Yu, 2006 [9] which defines consistency as the consistency of a model (LASSO) to select the true model. The true model in the present case is unknown and will remain unknown for the time being, so our aim is to describe the consistency of the model with regard to their robustness on slight changes of the penalty factor $\lambda$ since for each CV we are getting a slightly different optimal $\lambda$. 
2 Regression Models and Notations

In this section the models and notations used for the analyses are described. The overall notations used are $y \in \mathbb{R}^N$ as the dependant Variable / label, $X \in \mathbb{R}^{N \times D}$ as the independent / predictor variables, $\beta \in \mathbb{R}^D$ as regression coefficients and $\lambda \in \mathbb{R}$ as hyperparameter of the regularization. $\hat{\beta}$ denotes the resulting coefficients after a minimization process.

2.1 General Regression Approaches

Linear regression is a well known method to estimate values based in the manner of

$$y = X \beta$$  \hspace{1cm} (1)

In order to retrieve an optimal solution for this model $\beta$ has to be minimized. In a classic least-squares approach this leads to

$$\hat{\beta} = \min_{\beta} \|y - X\beta\|^2_2$$  \hspace{1cm} (2)

2.2 LASSO

The least absolute shrinkage and selection operator (LASSO) was introduced by Tibshirani, 1996 [10]. It is a regression method that minimizes the least-squares and has an additional penalty/regularization term for the regression coefficients given by the $L_1$-Norm.

$$\hat{\beta} = \min_{\beta} \|y - X\beta\|^2_2 + \lambda \|\beta\|_1$$  \hspace{1cm} (3)

Due to the $L_1$-Norm some of the coefficients are more likely set equal to zero, depending on the regularization parameter $\lambda$. Therefore it is not only possible to use the LASSO model for estimation (like in other regression methods), but also for variable selection.

In figure 1 you can see a geometric interpretation of the LASSO penalty term ($L_1$ norm). [3] You see on the left side, the estimation of the LASSO model in which the $\beta_2$ is set equal to zero due to the $L_1$-norm. The geometry of the $L_1$ norm is visualized by a scaled unit ball regarding the norm. On the right side one can see the same estimate visualized by a ridge regression method. In the case of the ridge regression estimate $\beta_2$ is small, but not set to zero.

For the minimization the glmnet [2] minimizer package for R is used. The solution path can be expressed by representing $\beta$ as a function of $\lambda$ (see Fig. 2). The glmnet package is providing logistic regression models, that is used for categorical variables.

Figure 1: Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error ans constraint functions. The solid blue areas are constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$ respectively, while the red ellipse are contours of the least squares error function. [3]

Figure 2: The Solutionpath for a single validation set is shown. The result is expected since with increasing $\lambda$ the penalty term $\lambda \|\beta\|_1$ is dominated by $\lambda$ and vice versa. Also the tendency of creating a sparse model by dismissing (setting to 0) coefficients due to the L1 norm can be observed.
2.3 Group LASSO

The Group LASSO Model is an extension of the LASSO model and separates the predictor set into \( J \) groups. \[7\] The corresponding minimization term is given by
\[
\min_{\beta} \|y - \sum_{j=1}^{J} X_j \beta_j \|_2^2 + \lambda \sum_{j=1}^{J} \|\beta_j\|_{K_j}
\]
with

Predictors: \( X_j \in \mathbb{R}^{N \times D_j} \), \( \forall j = 1, ..., J \)
Regression coefficients: \( \beta_j \in \mathbb{R}^{D_j} \), \( \forall j = 1, ..., J \)
\( K_j \in \mathbb{R}^{D_j \times D_j} \), positive definite \( \forall j = 1, ..., J \)
\( \|v\|_K = \sqrt{v^T K v} \)

Depending on the structure of the problem, each group can be penalized using different norms \( \|v\|_K \) induced by a scalar product (here represented by \( K \)). Different criteria for finding a suitable \( K \) can be found in Huang et al., 2012 \[5\]. For better comprehension and because we don’t have enough prior knowledge, about the structure of the data, we simply choose \( K \) as the Identity matrix. The choice of \( K = I \) is corresponding to the euclidean norm over the regression coefficients for each group \( \beta_j \).

The extension of Group LASSO for logistic regression is developed and already used for real world application i.e. splice cite detection in DNA sequences \[7\]. We used an implementation of the R package grpreg \[1\], it provides an implementation of Group LASSO for logistic regression.

2.4 Adaptive LASSO

The Adaptive LASSO Model \[13\] tries to achieve better prediction performance by introducing an elastic weighting on \( \beta \). Hui Zu et al. \[13\] show that for LASSO there exist certain criteria under which the consistency of LASSO to select the true model can be violated. Therefor an adapted model is introduced with
\[
\hat{\beta} = \min_{\beta} \|y - X \beta\|_2^2 + \lambda \sum_{i=1}^{D} w_i |\beta_i| \tag{4}
\]

The minimization process according to Hui Zu et al. \[13\] is calculated in the following manner:

**Algorithm 1 Minimization of Adaptive LASSO**

1. initial \( \beta^* \) estimation by running a crossvalidation with standard LASSO and selecting the model with \( \lambda = \lambda_{\text{min}} \) resulting in \( \beta^* = \hat{\beta}(\text{LASSO}) \)
2. The weighting is set to \( \hat{w} = \frac{1}{\hat{\beta}} \)
3. Setting \( X^* = \frac{X}{\hat{w}} \), a set of temporal coefficients \( \beta^{**} \) is obtained by
\[
\hat{\beta}^{**} = \min_{\beta} \|y - X^* \beta\|_2^2 + \lambda \sum_{i=1}^{D} |\beta_i|
\]
4. Finally we get \( \beta^* = \frac{\beta^{**}}{\hat{w}} \)

As we see this method includes several cross validations which makes the calculation quite heavy concerning calculation time. However the method is straight forward and since there already exists a reviewed implementation \[8\], a valid integration in our project was possible within the time scope.

2.5 Crossvalidation

The crossvalidation (CV) for standard LASSO utilizes the cv.glmnet implementation for R \[2\] which provides efficient minimization by pathwise coordinate descent for coefficient updates and a method called ”covariance update” which is a dynamic programming approach to increase the efficiency of the solver. Since our label is a binomial factor a model for logistic regression was needed. Fortunately this feature is also provided by the glmnet as well as the functionality to store and set the fold indices for the evaluation as well. This feature was necessary in order to to make sure that the set of folds is equal for all models for each run and to compare.

The CV for Group LASSO uses the grpreg R package \[1\], which has almost similar functionality regarding the logistic regression and supports CV.

For the CV with adaptive LASSO the parcor package \[8\], which has almost similar functionality regarding the logistic regression and supports CV.

For the CV with adaptive LASSO the parcor package was used \[8\]. However the included CV algorithm did not support binomial classifiers as well as the possibility to run the CV on specified folds. So some minor adaptions had to be done in order to make the CV work properly.
Figure 3: The misclassification error for the solution path of the crossvalidation shows a local minimum. $\hat{\beta}$ are the coefficients corresponding to the minimal value of $\lambda$ at this point.

3 Analysis

The idea is to generate a large set of optimal coefficients retrieved by multiple CVs of all models and to define an appropriate metric to describe the structure of the feature selection. Therefore we defined $M$ to be the number of CVs ('Runs') and $N$ to be the number of variables so that a coefficient vector is defined as $\beta_i^{(\alpha)}$ with $i \in [1, ..., M]$ and $\alpha \in [1, ..., N]$. Since we are only interested, if a coefficient was selected or not we define the number of selected coefficients as $V_i = \# \{ \alpha | \beta_i^{(\alpha)} \neq 0 \land \alpha = 1, ..., M \}$. In the following sections the term 'selected' refers to these coefficients whereby the term 'dismissed' denotes all coefficients that were set to 0.

The main metric to describe the selection behaviour is now defined as $\bar{V}_i = \frac{V_i}{M}$ which is in fact the mean of the selection indices and therefore called 'mean Selection'. However, since it is the mean of an index its interpretation differs from the normal mean of a distribution. It can be described as a distance criterion for selection persistence, which we define as selection consistency. As described in the introduction, it is noted that this consistency-term is not to be confused by the consistency of a model to select the true model.

The consistency criteria is defined by

$$C(\bar{V}) = \begin{cases} 1, & \text{if } \bar{V} < th \lor \bar{V} > 1 - th. \\ S(\bar{V}), & \text{otherwise.} \end{cases}$$

(5)

whereby $th$ is the threshold for which the selection is considered to be consistent and $S(\bar{V})$ is defined as an indicator for the strength of inconsistency:

$$S(\bar{V}) = \frac{|\bar{V} - 0.5|}{0.5 - th}$$

(6)

The reason behind this metric is if a variable was selected or dismissed in all runs, that variable is consistent selected. Whereas the maximum inconsistency would be if a variable is selected in half of the time and dismissed in the other half which is represented by $\bar{V}_i = 0.5$. Additionally a threshold $th$ is defined so that variables that are within the threshold they are considered as consistently selected as well (see Fig. 4).

Figure 4: The consistency strength $S$ is defined as continuous value between 0 and 1 that maps the distance of the mean Selection $\bar{V}_i$ according to its maximum uncertainty of 0.5.

3.1 Data

Raw Data

The data set we used is from a multi-centre study on smoking-related behavior [6]. The study participants were exposed to a battery of cognitive tests, self-reporting questionnaires as well as genetic testing and a 3 min EEG recording. The data set contains a set of means of a 3 minutes EEG recording and attributes of addiction of the probates in the form of a binomial factor. The EEG data is taken as predictor set contains 185 variables. Fig. 5 shows that the measurements from the different centers can be distinguished by their mean and max amplitude. Especially some centers seem to have fault measurements resulting in strong outliers. To reduce this effect on
the prediction we decided to pre-process the data and make a basic outlier detection.

Figure 5: The mean Value of all Observations of the predictor set.

Pre-processing
The Data was already z-standardized for each center so we only added a basic outlier detection depending on the mean and standard deviation (std) of each center with $C$ being the total number of centers.

$$X \in N_k \times D, k \in [1, ..., C]$$

$$X_t = \frac{1}{D} \sum_{i=1}^{D} X_{i,t}, t \in [1, ..., N_k]$$

$$s = \sqrt{\frac{1}{N_k} \sum_{i=1}^{N_k} (x_i - \bar{x})^2}$$

We defined a threshold of $3s$ so that all $x_i < 3s$ are accepted.

3.2 Experiment
At first a series with $M = 1000$ CVs for standard LASSO is executed using the `cv.glmnet` procedure. The necessary parameters are

- `nlambda=100` is the amount of tested lambda values.
- `family = "binomial"` is the option for logistic regression for the binomial label
- `type.measure = "class"` is the indicator for the evaluation method. "class" means that the misclassification Error is evaluation parameter.
- `alpha=1` is a hyperparameter that denotes the elastic-net mixing which could be used if a mixture of L1 and L2 penalty is wanted. alpha = 0 is used for ridge regression (L2) and alpha = 1 for pure LASSO regression.

To increase the reliability of the experiment, we manually set the machine seed for each set of sequential runs. The output is a 1 x N list of fit-objects (fit-objects are defined for the glmnet that can be directly used as input for `predict` in order to make predictions with the model) and a M x N Matrix containing all $\hat{\beta}$ and a list of FoldIDs.

Then the CVs for Adaptive Lasso and Group LASSO were calculated with the same input parameters and the foldID list. For adaptive LASSO no further parameters were necessary but for Group LASSO the additional input for the grouping of variables had to created.

For Group LASSO we have to specify the groups in advance. We separated the data by the location of the electrodes corresponding to the position of the electrodes used for the measurement. The electrodes are located on the scalp using the 10–20 system. The groups are separated based upon which area of the scalp the sites are located, additional to the direct measurements of a specific site we have also have the coherence between the electrodes. See Fig. 8.

Each group contains all corresponding frequency bands that where measured. A detailed description of the groups you can find in Tab. 6 and 7

3.3 Results
The results of the calculation of selection consistency $\tilde{V}_i$ for all models is plotted in Fig. 9 (for comparison you can find the $\tilde{V}_i$ for the individual groups in
<table>
<thead>
<tr>
<th>Group name</th>
<th>hemisphere</th>
<th>lobe</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>left</td>
<td>frontal</td>
</tr>
<tr>
<td>G2</td>
<td>right</td>
<td>frontal</td>
</tr>
<tr>
<td>G3</td>
<td>left</td>
<td>temporal</td>
</tr>
<tr>
<td>G4</td>
<td>right</td>
<td>temporal</td>
</tr>
<tr>
<td>G5</td>
<td>left</td>
<td>central</td>
</tr>
<tr>
<td>G6</td>
<td>right</td>
<td>central</td>
</tr>
<tr>
<td>G7</td>
<td>left</td>
<td>parietal</td>
</tr>
<tr>
<td>G8</td>
<td>right</td>
<td>parietal</td>
</tr>
<tr>
<td>G9</td>
<td>left</td>
<td>occipital</td>
</tr>
<tr>
<td>G10</td>
<td>right</td>
<td>occipital</td>
</tr>
<tr>
<td>G14</td>
<td>midline</td>
<td>F,C,P</td>
</tr>
</tbody>
</table>

Figure 6: Separation of the electrode measurements into groups based on the EEG electrode site based on the 10-20 system.

<table>
<thead>
<tr>
<th>Group name</th>
<th>hemisphere</th>
<th>sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>G11</td>
<td>left</td>
<td>F3, C3, T5, P3</td>
</tr>
<tr>
<td>G12</td>
<td>right</td>
<td>F4, C4, T4, P4</td>
</tr>
<tr>
<td>G13</td>
<td>between</td>
<td>F3-F4, C3-C4, P3-P4</td>
</tr>
</tbody>
</table>

Figure 7: Separation of the coherence between the electrode measurements into groups based on the EEG electrode site based on the 10-20 system. For G11 and G12 the groups contain the pairwise coherence between all listed sites. For G13 the electrode pairs are listed.

Figure 8: Separation of the groups depending on the location of the electrode sites (10–20 system). Groups based on measurement (left) and the coherence between electrodes (right).

Figure 9: Qualitative description of $\tilde{V}$ for for all models with consistency threshold $th$

Figure 10: Quantitative Analysis of Consistent and inconsistent features for all models and the respective mean Inconsistency strength $\bar{S} = 100 \times (1 - S)$

The analysis of mean and variance of the underlying data of consistent positive ($C^+$), consistent dismissed ($C^-$) and inconsistent ($I$) distributions did not deliver any interpretable results. To figure out the Appendix 19) and the quantitative description is displayed in Fig. 10. Concerning the amount of inconsistent betas as well as the strength of inconsistency the standard LASSO is located in between adaptive and Group LASSO. Group LASSO tends to select more variables but shows the least Inconsistency strength whereby adaptive LASSO shows a very conservative selection of features with only few inconsistent ones that are however strongly variant. Compared to standard LASSO both models both model try to overcome the uncertainty by being either very conservative (adaptive) or ”open” (Group) concerning feature detection.
if \( \tilde{V} \) or \( S \) can be linked to the underlying data we decided to analyze the Kernel Density and Quantile distribution. Figures 12, 13, 14 and fig. 15 for the overall distribution show the median, upper and lower quantile distribution for each model. The plots show the Quantile distribution for \( S^* \) which is defined as:

\[
S^* = \begin{cases} 
-1, & \text{if } C^+ . \\
0, & \text{if } C^- . \\
S, & \text{otherwise.}
\end{cases}
\]  

(8)

The Set of EEG Data is then ordered by \( S^* \) to have an impression of the selected features. We can see that LASSO and adaptive LASSO only vary little in their selection pattern with regard to the distribution. However the inconsistent \( S^* > 0 \) features tend to have a higher variance and are less structured than the dismissed variables. Group LASSO however seems to select features where the median and Quantile distribution has a very low variance.

The Kernel distributions for all configurations are found in the appendix. In Fig. 11 their main properties are described whereby 'Cluster' denotes the visual interpretation of clustered distributions. E.g. cluster = \([-0.3, 0.5]\) and \([1.2, 5]\) means that 2 separate distribution clusters within the written intervals could be estimated. In this set we are just looking at \( S = 1 \) for \((C^+ \cup C^-)\) and \( S < 1 \) for \((I)\) Selections. For standard LASSO (std) the Maximum density seems not to be related to \( S \) whereby adaptive (ada) and group (grp) LASSO show a notable difference in this concern. The description of clusters is quite difficult since it is just a visual description in the first place. However it roughly shows that std and grp have a notable shrinkage from \( S < 1 \) to \( S = 1 \) in their mean interval width from 1.3 and 2 to .55 to 1.03 respectively whereby ada stays more or less constant in this respect.

### 4 Conclusion

We successfully described techniques and new ideas to compare the Feature-Selection behavior of LASSO, Adaptive LASSO and Group LASSO. For evaluation we used empirical methods on real world data. It is to say that Adaptive LASSO had the best performance regarding the selection persistence on the data set we used, which we covered in detail in this paper. Group LASSO performed well regarding a low "Inconsistency Strength". Further investigation could be made by inspection of the performance of Group LASSO, when using other groups or compare the methods on other data sets. An other aspect, that was outside of the scope of this project, for comparison of these methods could be how well the methods perform regarding the model selection consistency described by Zhao et al. (2006) [12]. This would imply...
assumptions of a ground truth of the models with respect to the dataset which is only possible with more knowledge about the data. In the present case regarding the data distribution for consistent and inconsistent selections further investigations are necessary to draw actual conclusions from the presented descriptions.

Figure 13: Quantile distribution for Adaptive LASSO for $C^-, C^+$ and $I$ consistency

Figure 14: Quantile distribution for Group LASSO for $C^-, C^+$ and $I$ consistency

Figure 15: Mean EEG Distribution for all Models.
5 Appendix

(a) Adaptive Lasso Kernel Densities for $S = 1$

(b) Adaptive Lasso Kernel Densities for $S < 1$

(a) Group Lasso EEG Kernel Densities For $S = 1$

(b) Group Lasso EEG Kernel Densities For $S < 1$
Figure 19: The $\tilde{V}_i$ value for each group by index given by Tab. 6 and 7
References


