Model selection and estimation in regression with grouped variables

Clay Olsen[1]

[1]*Ming Yuan Georgia Institute of Technology, Atlanta, USA*
[2]*Yi Lin University of Wisconsin–Madison, US*

April 9, 2020

1 Topic/problem

Sometimes, we have variables that seem linked in structure. In these models, we can express these linked variables as a group of input variables. The most common example of this is with multi-factor analysis of variance (ANOVA) models, where a factor with several levels will be expressed through multiple dummy variables. We would aim to select the important variables and interactions necessary for accurate prediction. Also in additive models, we often have polynomial and non-parametric variables. In these models, we would select important groups of basis functions. In these situations, variable selection generally amounts to the selection of important factors, not just the individual input variables. It is useful to group input variables of the same factor for eventual analysis in these situations.

Traditional selection methods for these cases include best subset and stepwise procedures. The problem with these kinds of procedures is that they do not have a piece-wise linear solution path, so they are computationally expensive. Some methods such as lasso and least angle regression selection (LARS), have piece-wise linear solution paths, but are more tailored towards identifying individually important variables rather than groups of variables. This paper presents three methods tailored towards grouped variables selection: Group Lasso, Group LARS, and Group Non-negative Garrote (NNG). These methods treat variables in a grouped setting, while utilizing their computational efficiency and reliable accuracy under certain circumstances.

2 Previous model selection methods

The goal with model selection procedures is to simplify an equation, like the one below, down to only important predictors:

$$\mathrm{Y} = \sum_{i} X_j \beta_j + \epsilon$$

2.1 Best Subset and stepwise procedures

The best subset method utilizes a selection criterion, such as Akaike Information Criterion (AIC), to evaluate every candidate model. The "best subset" will be the subset of variables that minimizes the AIC. This method tends to be very computationally inefficient with even a moderate number of

variables because the number of candidate models increases exponentially with every addition of variables. So, for many cases, these is not a feasible model selection strategy.

Using a stepwise procedure, we can find a final model by adding variables to a null model and selecting the subset of variables that resulted in the lowest AIC. Or use backward selection method, where we start with a full model and drop variables until we have a null model. We then pick the model with the lowest AIC. While this method is more computationally efficient than the best subset method, it has restrictions of its own. Standard errors tend to be conservative, having bias towards zero. When parameters are far from 0, their standard error estimates will be too low and the confidence intervals will be too narrow. Due to this, stepwise tends to find locally optimal solution paths rather than globally optimal.

2.2 Lasso and Least Angle Regression Selection (LARS)

Lasso is a constrained version of Ordinary Least Squares, with its added penalty function. The general method has the following steps: 1) A solution path is indexed by a tuning parameter (λ) is built, 2) a final model is selected on that tuning path determined by cross validation or a criterion like AIC/BIC. Lasso is piece-wise linear, so it is computationally inexpensive compared to methods like best subset.

LARS is similar to the step-wise forward selection method. LARS starts with the null model and adds the predictor variable with the most correlation with the response variable. It will increase the coefficient in the sign of its correlation until another variable has as much correlation with the response. It continues this method taking steps and picking the final model based on minimizing cross-validation error or a criterion like AIC or BIC.

These methods are popular because of their piece-wise linear solution paths. Both methods are great for selecting individual input variables, but not for selecting grouped variables. Both methods selects variables on individual strength rather than the group strength. Since these methods treat grouped variables individually, they also do poorly when a model has high multicollinearity.

3 Group Variable Selection Methods

While efficient, the available piece-wise linear solution pathed methods perform poorly with grouped variables. The paper presents modified versions of these methods, specialized for grouped variables called Group Lasso, Group LARS, and Group Non-negative Garrote (Group NNG).

3.1 Group Lasso

$$\beta_{\text{grLASSO}}(\lambda) = \arg\min \|Y - \sum_j X_j \beta_j\|^2 + \lambda \sum_j \|\beta_j\|_k$$

$$\|\beta_l\|_2 = \sqrt{\beta_1^2 + \beta_2^2 + \beta_3^2 \ldots + \beta_p} \text{(for group l, with p predictors in the group)}$$

Group Lasso is very similar to Lasso, except for the penalty function in Group Lasso takes the magnitude of the coefficients of a group, rather than the penalty function Lasso uses, which takes the absolute value of individual coefficients (l1-norm). Group lasso will shrink and select variables of a group together. Using the Group Lasso penalty with grouped variables encourages sparsity at a factor level, which prevents against over-fitting the data. The Group Lasso algorithm is generally very stable and reaches a convergence tolerance within a few iterations. However, the function

becomes much more computationally expensive with an increase in predictor variable groups. This is due to the fact that Group Lasso is not piece-wise linear except under certain conditions:

Theorem: The solution path of the group lasso is piece-wise linear if and only if any group lasso solution β can be written as $\beta_j = c_j\beta_j$, j= 1,. .., J, for some scalars cl,..., cJ.

This means, the Group Lasso function is only piece-wise linear if each group only has one predictor, or if the data is orthonormal. This is a considerably tight restriction to maintain a piece-wise linear solution path and rarely holds true in practice.

3.2 Group Least Angle Regression Selection

The LARS algorithm follows a similar methodology to forward selection. Starting with the null model (all coefficients = 0), the LARS algorithm finds the predictor variable most correlated with the response variable and proceeds adding variables in this direction. The algorithm will increase the coefficient of the first variable (x_1) until another variable (x_2) has as much correlation with the response variable as x_1 has with the residual. The algorithm will proceed in the joint least squares direction of x_1 and x_2 until a third variable is of equal correlation and joins the set. LARS will continue in the direction that has an equal angle to the three variables until a fourth enters. Lars will continue this process until all variables have joined the model. Then, the model with the lowest AIC is picked. The solution path for LARS is piece-wise linear, so it is computationally efficient even with a large number of variables.

3.3 Group Non-negative Garrotte (Group NNG)

$$\beta_{\text{grNNG}}(\lambda) = \arg\min\|\text{Y} - \sum_j X_j\beta_j\|^2 + \lambda\sum_j d_j$$

For group Non-negative Garrotte, the coefficient estimation, β_j, is the least squares estimate, scaled by a non-negative constant d_j, rather than the magnitude of coefficients we saw in Group Lasso. Group NNG requires the solution to the least squares and does not work if the sample size, n, is less than the number of variables, p. The solution path for Group NNG is piece-wise linear and, therefore, computationally inexpensive.

4 Comparison of Group Selection Methods

While the three models are tailored to work with grouped variables, they have restrictions that inhibit their performance. Group Lasso has trouble dealing with highly correlated variables in different groups. Group Non-negative Garrote is reliant on solving the least squares of the model and tend to struggle with models that have a high number of variables relative to sample size. If the number of variables is greater than the sample size, we can not use Group NNG. These restrictions make these methods sub-optimal with some data sets.

As discussed in the previous section, Group LARS and Group NNG maintain a piece-wise linear solution path, while Group Lasso only has a piece-wise linear solution path if every group has only one input variable or if the data is orthonormal. In general, Group Lasso will be more computationally expensive than Group LARS and Group NNG in large scale problems.

5 Group Model Section Methods with Data

5.1 Baby Weight Data

To compare Group Lasso, Group LARS, and Group NNG, we examine the Birth Weight Data set from, "Hosmer and Lemshow 1989." The Birth Weigh data set contains the records of 189 birth weights and eight predictors concerning the mother.

Explanatory Variables

1. Mother age: (continuous) added to 3rd order polynomial
2. Mothers Weight: (continuous) in lbs, to the 3rd order polynomial environment.
3. Race: (factor) Black, White, Other
4. Smoker: (factor) Yes, No
5. Number of premature labors: (factor) 0, 1 , 2
6. History of Hypertension: (factor) yes, no
7. Presence of Uterine Irritability: (factor) yes, no
8. Number of physician visits during first trimester: (factor) 0, 1, 2

Response

1. Baby Weight: (continuous) grams

5.2 Group Methods and Lasso Final Models

For comparison, I also used the regular Lasso method for variable selection. Below are the plots showing the methods coefficient values for tuning parameter for the Lasso, Group Lasso, and Group NNG plots. The Group LARS plot shows the models BIC for every step. The plots are preceded by the final models resulting from each model selection method.

Table 1: Lasso Model

(Intercept)	age1	age2	age3	lwt1	lwt2	lwt3	white	black	smoke
3.000	0	0.403	0.420	0	0	0.062	0.173	0	–0.197
ptl1	ptl2m	ht	ui	ftv1	ftv2	ftv3m			
0	0	–0.294	0.007	0	0	–0.031			

Table 2: Group Lasso Model

(Intercept)	age1	age2	age3	lwt1	lwt2	lwt3	white	black	smoke
3.043	0	0	0	0	0	0	0.140	–0.044	–0.224
ptl1	ptl2m	ht	ui	ftv1	ftv2	ftv3m			
0	0	–0.021	–0.347	0	0	0			

Table 3: Group Least Angle Regression Selection Model

(Intercept)	white	smoke	ui
3.005	0.392	−0.426	−0.472

Table 4: Group Non-negative Garrotte Model

(Intercept)	age1	age2	age3	lwt1	lwt2	lwt3	white	black
3.065	0	0	0	0	0	0	0.146	−0.043
smoke	ptl1	ptl2m	ht	ui	ftv1	ftv2	ftv3m	
−0.261	0	0	−0.025	−0.413	0	0	0	

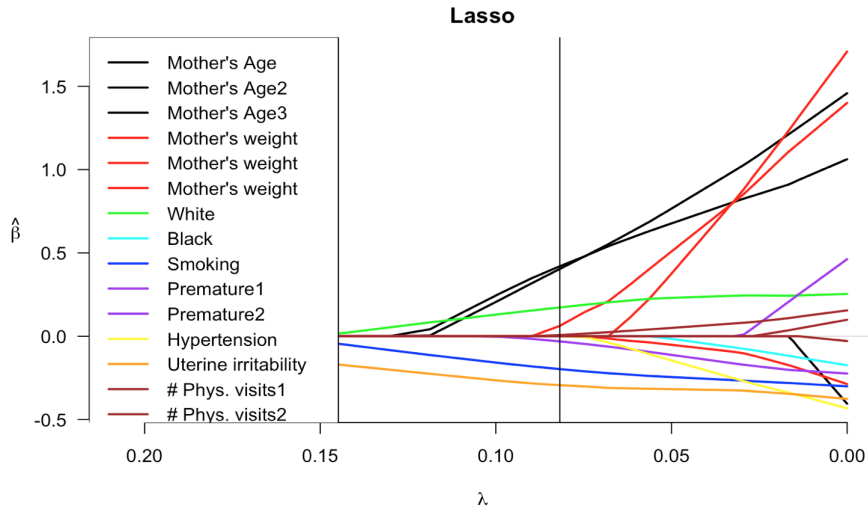

Figure 1: Model selection for Lasso by tuning parameter

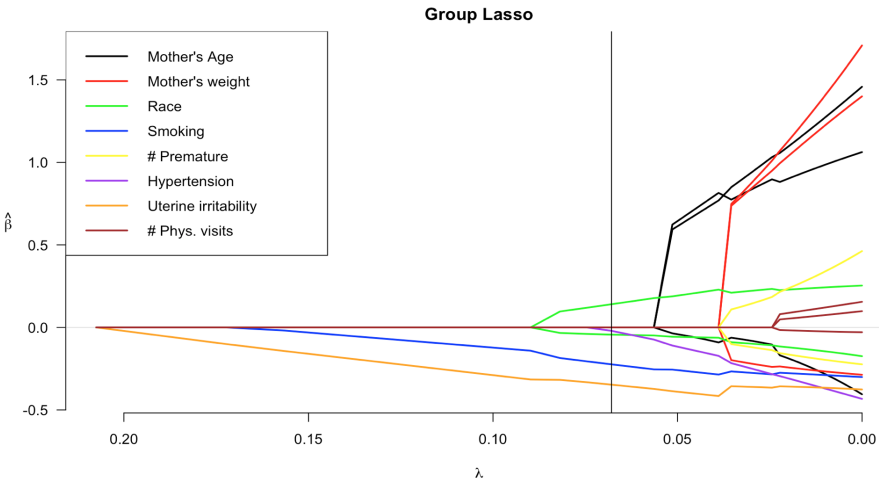

Figure 2: Model selection for Group Lasso by tuning parameter

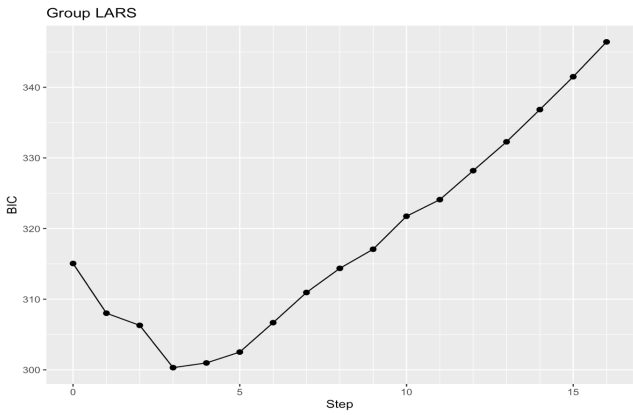

Figure 3: Model selection for Group LARS by BIC criterion

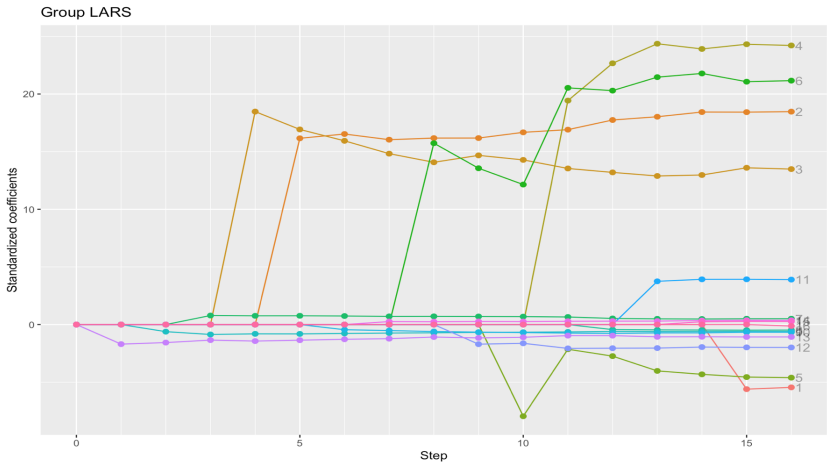

Figure 4: Coefficients for LARS by step

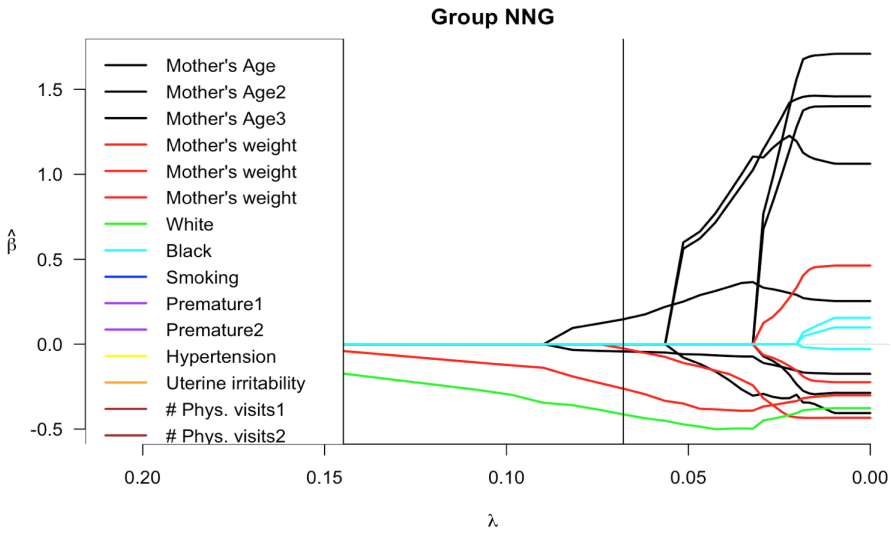

Figure 5: Model selection for Group NNG by tuning parameter

Looking at the Lasso model and Group Lasso model, we can see a difference in how the method turns coefficients to 0 with groups defined. In the Lasso model, we can see the coefficients of grouped variables such as Mother's Weight, Mother's Weight^2, and Mother's Weight^3 converge to zero at different tuning parameters. In the final model for Lasso, Mother's Weight^3 remained in the model while the other variables in the group were zeroed. On the other hand, with Group Lasso, variables in the same group are zeroed at the same tuning parameter level. Between these two methods, we can clearly see the impact of grouping predefined "grouped" variables. The Group NNG model shared a similar solution path with the Group Lasso model.

Looking at the Group LARS model, we can see the forward selection process. Graph (a) shows the BIC for every step or addition of a variablele, and we can see that the optimal model based of BIC criterion would be the model built step 3. In this case, the Group LARS model removed the most variables of the model selection methods.

5.3 Prediction Accuracy

methods <fctr>	Prediction error <dbl>
prediction_las	141.5344
prediction_grlas	141.8166
predict_grlars	142.7232
predict_grNNG	142.8267

Figure 6: Model selection methods prediction Accuracy

For validation, the data set was split into a training set with three-quarters of the data and a test set with the remaining quarter like it was done in the paper. All of the methods resulted in number of premature labors being dropped from the final model. The group methods, additionally, dropped age, mother's weight, and visits to a physician during the first trimester. Looking at the prediction table we can see very little difference between these modeling methods. This is likely due to the fact that the Baby Weight data set only has 189 observations and only 8 factors. A larger and more diverse data set with varying leveled factors may show greater results with the grouping method.

6 Conclusion

Group Lasso, Group LARS, and Group NNG are suitable counterparts for their non-grouped alternatives when working with grouped data. While the three methods performed similarly with this example data set, there are trade offs for each method.

As previously discussed, Group Lasso is not piece-wise linear unless all factor levels are one or the data is orthonormal. In practice, this is very unlikely to occur. Despite Group Lasso's accuracy, it tends to be very computationally expensive in large scale problems. Additionally, Group Lasso struggles with a lot of auto-correlation between variable groups. Group LARS has comparable accuracy to Group Lasso and can be computed quickly due to its piece-wise linear solution path. Group NNG had the fastest computing time of the three methods and similar accuracy, but suffers from some restrictions. Group NNG requires the least squares solution and is sensitive to the assumptions of ordinary least squares. Additionally, the accuracy of Group NNG diminishes when the number of variables is high relative to the sample size. If the number of predictors is larger than the sample size, Group NNG cannot be applied.

6.1 Limitations

The data set for this paper was fairly limiting in producing results that showed clear differences with model selection method. The paper did a simulation study where they were able to create data sets that would greater show the differences in performance for the different methods, especially compared to their non-grouped counterparts. A real data set may be more clear for representing the differences of the models. For example, with a set that has a high number of predictors relative to the sample size, we would expect a poor performance with the Group NNG method compared to its counterparts. Additionally, the Group NNG method is not very popular and had no available R package. Further efforts into implementing Group NNG could be taken to see if there are scenarios where it is a worthwhile method to use.

References

[1] Yuan, Ming, and Yi Lin. “Model Selection and Estimation in Regression with Grouped Variables.” Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, 2006, pp. 49–67., doi:10.1111/j.1467-9868.2005.00532.x.

[2] "Statistical Learning and Data Mining Codes" (n.d.). Available at http://www.biostat.umn.edu/ weip/course/dm/examples/exampleforhighd1.R

[3] "Package robustHD" https://cran.r-project.org/web/packages/robustHD/robustHD.pdf

[4] Hosmer, D. W., and S. Lemeshow. Applied Logistic Regression.: Wiley Series in Probability and Mathematical Statistics. Wiley, 1989.