Study of "Model Selection and Estimation in Regression with Grouped Variables"

STOR 767 Reading Project

Department of Statistics and Operations Research

UNC Chapel Hill

Report Scribes: Leo Li (PID: 730031954)

October 12, 2020

1 Summary

The paper "Model Selection and Estimation in Regression with Grouped Variables"[1] is about the problem of selection of grouped variables for accurate prediction in regression. The problem of group selection arises in many practical settings in which there are natural group structures among the covariates. Given the limitation of some of the existing methods such as multifactor ANOVA and stepwise backward elimination, Yuan and Lin focus on the accuracy of estimation and consider extensions of LASSO, the LARS algorithm, and the non-negative garrotte for group selection. In particular, Yuan and Lin have extended LASSO, the LARS algorithm, and the non-negative garrotte to the setting of group selection, proposed efficient algorithms for computation, and also showed that these extensions give superior performance to the existing methods. They also study the similarities and differences among these methods. Simulation studies and real data analysis are also used to provide further evidence for the proposed conclusion.

2 Problem

In this paper, the main research problem is group selection, which is different from variable selection in the sense that, the goal is to select some groups of variables, instead of single variables, that are associated with outcomes. This problem is important and worthy of attention because in some of the practical settings, there are natural group structures among the covariates (e.g., genes and pathways), and it would be helpful for us to consider the covariates at the level of groups. The paper "Model Selection and Estimation in Regression with Grouped Variables" proposes and studies several methods that produce accurate prediction while selecting a subset of important groups. Specifically, by extending LASSO, LARS, and non-negative garrotte, Yuan and Lin proposes group LASSO, group LARS, and group non-negative garrotte to conduct group selections.

3 Motivations

New research on this problem is needed because there are few existing methods that can be used to conduct group selection, especially when there are a lot of groups of predictors that need to be considered. In addition, there are various drawbacks of the existing methods, which will be described in the following paragraphs.

One of the existing methods is the multifactor ANOVA, in which each factor may have several levels and can be expressed through a group of dummy variables. The goal of ANOVA is often to select important main effects and interactions for accurate prediction, which amounts to the selection of groups of derived input variables. In the case of multifactor ANOVA models with balanced design, it is possible to construct an ANOVA table for hypothesis testing by partitioning the sums of squares. The columns in the full design matrix X are orthogonal; thus the test results are independent of the order in which the hypotheses are tested. However, there is one drawback of ANOVA that, in the case of unbalanced design, in which the columns of X are no longer orthogonal, and there is no unique partition of the sums of squares so that the test result on one factor will depend on whether other factors are present or absent.

Other existing methods also include some traditional approaches to model selection, such as the best subset selection and stepwise procedures. In best subset selection, an estimation accuracy criterion, such as AIC or C_p, is evaluated on each candidate model and the model that is associate with the smallest score is selected as the best model. One drawback of this method is that the number of candidate models grows exponentially as the number of groups increases, so that even moderate numbers of groups of predictors seem impractical. In addition, although the stepwise procedures are computationally less expensive, such methods often lead to locally optimal solutions rather than the globally optimal solution.

4 Novel insights

The most novel and brilliant insights that lead to the proposed approach is that the authors make use of some of the recently proposed methods for variable selection, and then extend these methods to the setting of group selection.

Specifically, we can express the setting of group selection as,

$$Y = \sum_{j=1}^{J} X_j \beta_j + \epsilon,$$

where Y is an $n \times 1$ vector, $\epsilon \sim N_n(0, \sigma^2 I)$, X_j is an $n \times p_j$ matrix corresponding to the j^{th} group and β_j is a coefficient vector of size p_j, $j = 1, \ldots, J$. The problem of variable selection, which has been well-studied, can be regarded as a special case of group selection, in which $p_1 = \ldots = p_J = 1$. Therefore, one potential way to develop new methods for group selection is to consider extending and generalizing some existing variable selection methods that have been proved to have nice properties, and see if they can be adapted for the setting of group selection.

During the recent years, there has been a number of variable selection methods introduced. First, Tibshirani (1996)[2] proposed least absolute shrinkage and selection operator (LASSO), which is defined as,

$$\hat{\beta}^{LASSO}(\lambda) = argmin_\beta(\|Y - X\beta\|^2 + \lambda\|\beta\|_{l_1}),$$

where λ is a tuning parameter and $\|.\|_{l_1}$ is the l_1-norm, which induces sparsity in the solution.

The second variable selection method is the least angle regression selection (LARS) proposed by Efron *et. al.* (2004)[3], and it has also been shown that LARS is closely related to LASSO. The LARS algorithm can be described as follows: LARS algorithm starts with all the regression coefficients equal to 0 and finds the input variable that is most correlated with the response variable and proceeds in this direction. Instead of taking a full step towards the projection of Y on the variable, the LARS algorithm takes only the largest step that is possible in this direction until some other input variable has as much correlation with the current residual. At this point, the projection of the current residual on the space that is spanned by the two variables has an equal angle with the two variables, and the LARS algorithm proceeds in this direction until a third variable 'earns its way into the most correlated set'. The LARS algorithm then proceeds in the direction of the projection of the current residual on the space that is spanned by the three variables, a direction that has an equal angle with the three input variables, until a fourth variable enters, etc. The great computational advantage of the LARS algorithm comes from the fact that the LARS path is piecewise linear.

Another method for variable selection is the non-negative garrotte porposed by Breiman (1995)[4], and its estimate of β_j is the least square estimate $\hat{\beta}_j^{LS}$ scaled by a constant $d_j(\lambda)$. The shrinkage factor $d(\lambda) = (d_1(\lambda), \ldots, d_p(\lambda))'$ is given as the minimizer to

$$\frac{1}{2}\|Y - Zd\|^2 + \lambda \sum_{j=1}^{p} d_j,$$

subject to $d_j \geq 0$ for all j, and $Z = (Z_1, \ldots, Z_p)$, $Z_j = X_j \hat{\beta}_j^{LS}$, and $\lambda > 0$ is a tuning parameter. Then the non-negative garrotte estimate of the regression coefficient is $\hat{\beta}_j^{NG}(\lambda) = d_j(\lambda)\hat{\beta}_j^{LS}$, $j = 1, \ldots, p$.

The novel insight of this paper is that we can extend these 3 methods of variable selection (LASSO, LARS, and non-negative garrotte) to the setting of group selection. In "Model Selection and Estimation in Regression with Grouped Variables", Yuan and Lin present that these natural extensions of LASSO and LARS, called group LASSO and group LARS, respectively, improve over the LASSO and LARS in terms of group selection and enjoy superior performances to that of existing methods for group selection. Additionally, the relationship between the group LASSO and group LARS has also been studied, and this paper presents that these two methods are equivalent when the full design matrix X is orthogonal, but can be different in more general situations. A somewhat surprising result is that the solution path of the group LASSO is generally not piecewise linear whereas the solution path of group LARS is. Also considered is a group version of the non-negative garrotte. These three methods have been compared in the paper using simulations studies, and an easily computable C_p-criterion is also proposed to select the final model of the group selection.

5 The Method

The first method proposed in this paper is group LASSO. We start by introducing some notations. For a vector $\eta \in \mathbf{R}^d$, $d \geq 1$, and a symmetric $d \times d$ positive definite matrix K, we denote that,

$$\|\eta\|_K = (\eta' K \eta)^{1/2}.$$

We also denote that $\|\eta\| = \|\eta\|_{I_d}$. Given positive definite matrices $K_1, \ldots, K_J$, the group LASSO estimate is defined as the solution to

$$\frac{1}{2}\|Y - \sum_{j=1}^{J} X_j \beta_j\|^2 + \lambda \sum_{j=1}^{J} \|\beta_j\|_{K_j},$$

where $\lambda \geq 0$ is a tuning parameter. A common choice of K_j is that $K_j = p_j I_{p_j}$, for $j = 1, \ldots, J$. Then we can observe that, the objective function of group LASSO is constructed by a quadratic loss and a group LASSO penalty. The group LASSO penalty is an intermediate between the l_1-penalty that is used in the LASSO and the l_2-penalty that is used in ridge regression. Specifically, at the group level, it is l_1-penalty, penalizing the magnitude of a group. The groups with small magnitude will be shrunken to 0, and thus it enables group selection. Within the group, it is l_2-penalty, which cannot conduct within-group selections. The optimization algorithm is blockwise gradient descent algorithm, which is also an extension of gradient descent algorithm used to solve LASSO problem. Each iteration uses block soft-thresholding to provide a closed form solution. To be more specific, if $K_j = p_j I_{p_j}$, $j = 1, \ldots, J$, then a necessary and sufficient condition for $\beta = (\beta_1', \ldots, \beta_J')'$ to be a solution to group LASSO is that,

$$-X_j'(Y - X\beta) + \frac{\lambda \beta_j \sqrt{p_j}}{\|\beta_j\|} = 0, \forall \beta_j \neq 0,$$

$$\| - X_j'(Y - X\beta)\| \leq \lambda \sqrt{p_j}, \forall \beta_j = 0.$$

Since $X_j' X_j = I_{p_j}$ by standardization, it can be verified that the solution to the above equations is that,

$$\beta_j = (1 - \frac{\lambda \sqrt{p_j}}{\|S_j\|})_+ S_j,$$

where $S_j = X_j'(Y - X\beta_{-j})$, with $\beta_{-j} = (\beta_1', \ldots, \beta_{j-1}', 0', \beta_{j+1}', \ldots, \beta_J')$. The solution to group LASSO can then by obtained iteratively by applying the equation to $j = 1, \ldots, J$.

The second method proposed in this paper is group LARS, which is just a natural extension of LARS to the setting of group selection. We start by considering a simplified setting of group selection where $p_1 = \ldots = p_J = p$, and p does not necessarily equal to 1. Define the angle $\theta(r, X_j)$ between an n-vector r and a factor that is represented by X_j as the angle between the vector r and the space that is spanned by the column vectors of X_j. Therefore, $\cos^2\{\theta(r, X_j)\}$ is the proportion of the total variance of squares in r that is explained by the regression on X_j, so that we have,

$$\cos^2\{\theta(r, X_j)\} = \frac{\|X_j' r\|^2}{\|r\|^2}.$$

Then, following the same spirit of LARS, the group LARS would start with all coefficient vectors equal to the zero vector, and finds the factor, say X_j, that has the smallest angle with Y (i.e., $\|X_j' Y\|^2$ the largest) and proceeds in the direction of the projection of Y on the space that is spanned by the factor until some other factor has as small an angle with the current residual (i.e., the projection of the current residual on the space spanned by the columns of the design matrices of the two groups has equal angle with the two factors), and group LARS proceeds in this direction. As group LARS marches on, the direction of projection of the residual on the space that is spanned by the two groups does not change. Group LARS continues in this direction until a third factor has the same angle with the current residual as the two factors with the current residual. Group LARS then proceeds in the direction of the projection of the current residual on the space that is spanned by the three factors, a direction that has equal angle with the three factors, until a fourth factor enters, etc. In order to adjust for the setting in which different groups can have different predictors, a scaling factor of p_j, the size of the coefficient vector for the j^{th} group, can be introduced, so that we would like to find the group with largest $\|X_j'Y\|^2/p_j$, instead of $\|X_j'Y\|^2$. Then we would summarize the algorithm for group LARS as the following:

- Start from $\beta^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$.
- Compute the current 'most correlated set', $\mathcal{A}_1 = argmax_j \|X_j' r^{[k-1]}\|^2/p_j$.
- Compute the current direction γ which is a $p = \Sigma p_j$ dimensional vector with $\gamma_{\mathcal{A}_k^c} = 0$, and $\gamma_{\mathcal{A}_k} = (X_{\mathcal{A}_k}' X_{\mathcal{A}_k})^- X_{\mathcal{A}_k}' r^{[k-1]}$, where $X_{\mathcal{A}_k}$ denotes the matrix comprised of the columns of X corresponding to $\mathcal{A}_k$.
- For every $j \notin \mathcal{A}_k$, compute how far the group LARS algorithm will progress in direction γ before X_j enters the most correlated set. This can be measured by an $\alpha_j \in [0, 1]$ such that $\|X_j'(r^{[k-1]} - \alpha_j X\gamma)\|^2/p_j = \|X_{j'}'(r^{[k-1]} - \alpha_j X\gamma)\|^2/p_{j'}$, where j' is arbitrarily chosen from $\mathcal{A}_k$.
- If $\mathcal{A}_k \neq \{1, \ldots, J\}$, let $\alpha = \min_{j \notin \mathcal{A}_k}(\alpha_j) \equiv \alpha_{j^*}$ and update $\mathcal{A}_{k+1} = \mathcal{A} \cup \{j^*\}$; otherwise, set $\alpha = 1$.
- Update $\beta^{[k]} = \beta^{[k-1]} + \alpha\gamma$, $r^{[k]} = Y - X\beta^{[k]}$ and $k = k + 1$. Go back to the third bullet until $\alpha = 1$.

The last method proposed in this paper is group non-negative garrotte, which is naturally extended to the setting of group selection. Under the setting of group selection, $\hat{\beta}_j^{LS}$ is a vector, and every component of vector $\hat{\beta}_j^{LS}$ by the same constant $d_j(\lambda)$. To take into account the different number of derived variables in the factor, we define $d(\lambda)$ as,

$$d(\lambda) = argmin_d(\frac{1}{2}\|Y - Zd\|^2 + \lambda \sum_{j=1}^{J} p_j d_j),$$

subject to $d_j \geq 0$, for all j. The algorithm is pretty similar to group LARS, with a complicating factor of non-negativity constraints in the equation above. The algorithm can be described as follows:

- Start from $d^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$.
- Compute the current active set $\mathcal{C}_1 = argmax_j(Z_j' r^{[k-1]}/p_j)$.
- Compute the current direction γ, which is a p-dimensional vector defined by $\gamma_{\mathcal{C}_k^c} = 0$ and $\gamma_{\mathcal{C}_k} = (Z_{\mathcal{C}_k}' Z_{\mathcal{C}_k})^- Z_{\mathcal{C}_k}' r^{[k-1]}$.
- For every $j \notin \mathcal{C}_k$, compute how far the group non-negative garrotte will progress in direction γ before X_j enters the active set. This can be measured by an α_j such that $Z_j'(r^{[k-1]} - \alpha_j Z\gamma)/p_j = Z_{j'}'(r^{[k-1]} - \alpha_j Z\gamma)/p_{j'}$ where j' is arbitrarily chosen from C_k.
- For every $j \in \mathcal{C}_k$, compute $\alpha_j = \min(\beta_j, 1)$ where $\beta_j = -d_j^{[k-1]}/\gamma_j$, if non-negative, measures how far the group non-negative garrotte will progress before d_j becomes 0.
- If $\alpha_j \leq 0$, $\forall j$, or $\min_{j:\alpha_j>0}\{\alpha_j\} > 1$, set $\alpha = 1$; otherwise, denote $\alpha = \min_{j:\alpha_j>0}\{\alpha_j\} \equiv \alpha_{j^*}$. Set $d^{[k]} = d^{[k-1]} + \alpha\gamma$. If $j^* \notin \mathcal{C}_k$, update $\mathcal{C}_{k+1} = \mathcal{C}_k \cup \{j^*\}$; otherwise update $\mathcal{C}_{k+1} = \mathcal{C}_k - \{j^*\}$.
- Set $r^{[k]} = Y - Zd^{[k]}$ and $k = k + 1$. Go back to the third bullet until $\alpha = 1$.

Finally, Yuan and Lin also introduce a simple approximate C_p-type criterion to select the final estimates, which is based on the fact that, in Gaussian regression problems, for an estimate $\hat{\mu}$ of $\mu = E(Y|X)$, an unbiased estimate of the true risk $E(\|\hat{\mu} - \mu\|^2/\sigma^2)$ is,

$$C_p(\hat{\mu}) = \frac{\|Y - \hat{\mu}\|^2}{\sigma^2} - n + 2df_{\mu,\sigma^2},$$

where,

$$df_{\mu,\sigma^2} = \sum_{i=1}^{n} Cov(\hat{\mu}_i, Y_i)/\sigma^2,$$

which consists of unknown values. Yuan and Lin has proposed the following approximations under different settings: for group LASSO,

$$\tilde{df}\{\hat{\mu}(\lambda) \equiv X\beta\} = \sum_j I(\|\beta_j\| > 0) + \sum_j \frac{\|\beta_j\|}{\|\beta_j^{LS}\|}(p_j - 1);$$

for group LARS,

$$\tilde{df}(\hat{\mu} \equiv X\beta^{[k]}) = \sum_j I(\|\beta_j^{[k]}\| > 0) + \sum_j (\frac{\sum_{l<k} \|\beta_j^{[l+1]} - \beta_j^{[l]}\|}{\sum_{l<J} \|\beta_j^{l+1} - \beta_j^{[l]}\|})(p_j - 1);$$

and for non-negative garrotte,

$$\tilde{df}\{\hat{\mu}(\lambda) \equiv Zd\} = 2\sum_j I(d_j > 0) + \sum_j d_j(p_j - 2).$$

When the design matrix X is orthonormal, all of the above approximations are unbiased, and the performance of this approximate C_p-criterion is generally comparable with that of fivefold cross-validation, but the fivefold cross-validation is much more computationally expensive.

6 Applications

There are several advantages of the proposed methods. Specifically, all of group LASSO, group LARS, and group non-negative garrotte can be used in ANOVA problems with the general design and tend to outperform the traditional stepwise backward elimination method. The group LASSO enjoys excellent performance but its solution path is generally not piecewise linear and therefore requires intensive computation in large-scale problems. The group LARS method has comparable performance with that of the group LASSO and can be computed quickly owing to its piecewise linear solution path. The group non-negative garrotte can be computed the fastest among the three methods proposed, through a new algorithm taking advantage of the piecewise linearity of its solution.

However, there are also some limitations to the three methods proposed. Specifically, the methods proposed in this paper are only applicable in the setting of "all-in-all-out" selection (i.e., once one member in a group is selected, all other members will also be selected), but they cannot be used to further select which members of the group are important. In the case that the within-group selection is of interest, other approached should be used. An example would be the group bridge as proposes by Huang *et. al.* (2009)[5]. Furthermore, there is also a limitation specifically for group non-negative garrotte, owing to its explicit dependence on the full least squares estimates, in problems where the sample size is small relative to the total number of variables, the non-negative garrotte may perform suboptimally. In particular, the non-negative garrotte cannot be directly applied to problems where the total number of variables exceeds the sample size, whereas group LASSO and group LARS can.

There is one application that Yuan and Lin presented in the paper about the association between the birth weight of 189 babies and various predictors including mother's demographics, medical history, and many others. The data is split into a training data set which consists of three-quarters of observations and a testing data set which consists of one-quarters of observations. The training data set is used for model fitting and testing data set for validation. The paper presents the results such that the prediction errors for group LASSO, group LARS, and group non-negative garrotte are smaller than the stepwise procedure, which further supports the paper's conclusion such that the group LASSO, group LARS, and group non-negative garrotte can be used for selection of grouped variables for accurate prediction in regression.

References

[1] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49-67.

[2] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

[3] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics, 32*(2), 407-499.

[4] Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics, 37*(4), 373-384.

[5] Huang, J., Ma, S., Xie, H., & Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika, 96*(2), 339-355.