Model Selection and Estimation in Regression with Grouped Variables

Leo Li

University of North Carolina at Chapel Hill

Department of Statistics and Operations Research

Research Questions

- The main research problem is **group selection**.
- More specifically, we consider the problem of selection of some groups of variables for accurate prediction in regression.
- Consider some practical settings in which there are natural group structures among the covariates (e.g. genes and pathways, etc.).
- **Aim:** Select groups of variables that are associated with the response (All-in-all-out selection: once one member in a group is selected, all other members will also be selected).

Research Questions (Continued)

- We can express the setting of group selection as that,

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

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, $j = 1, \ldots, J$.

- Group selection 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.
- The problem of variable selection can be regarded as a special case of group selection, in which $p_1 = \ldots = p_J = 1$.

Existing Method 1: Multifactor ANOVA

- 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, and thus the test results are independent of the order in which the hypotheses are tested.
- **Drawback:** in the case of unbalanced design, 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.

Existing Method 2: Other Traditional Approaches

The Best Subset Selection

- 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.
- **Drawback:** the number of candidate models grows exponentially as the number of groups increases, so that even moderate numbers of groups seem impractical.

Stepwise Procedures

- The stepwise procedures are computationally less expensive.
- **Drawback:** such methods often lead to locally optimal solutions rather than globally optimal solution.

Novel Insights

- **Key idea:** make use of some of the recently proposed methods for variable selection, and then extend these methods to the setting of group selection.
- Generalize LASSO to group LASSO; generalize LARS to group LARS; and generalize non-negative garrotte to group non-negative garrotte.
- **Results:** group LASSO, group LARS, and group non-negative garrotte enjoy superior performance to that of existing methods for group selection.

Variable Selection Methods Revisited: LASSO

- LASSO (Least Absolute Shrinkage and Selection Operator) was proposed by Tibshirani (1996).
- LASSO problem 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.
- LASSO problem can be solved by coordinate gradient descent algorithm (essentially to solve the optimization problem one coordinate at a time, and at each step, the problem has a closed form solution).

Variable Selection Methods Revisited: LARS

- LARS (Least Angle Regression Selection) was proposed by Efron *et. al.* (1996).
- LARS Algorithm
  1. Standardize the predictors to have mean zero and unit norm. start with residual $r = y - \bar{y}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
  2. Find the predictor x_j most correlated with r.
  3. Move β_j from 0 towards its least-square coefficient $\langle x_j, r \rangle$, until some other competitor x_k has as much correlation with the current residual as does x_j.
  4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on (x_j, x_k), until some other competitor x_l has as much correlation with the current residual.
- The great computational advantage of the LARS algorithm comes from the fact that the LARS path is piecewise linear.

Variable Selection Methods Revisited: Non-negative Garrotte

- Non-negative garrotte was proposed by Breiman (1995).
- It is a variable selection method that shrinks the least squares estimates and puts some of these coefficients equal to zero.
- The estimate of β_j is the least square estimate $\hat{\beta}_j^{LS}$ scaled by a constant $d_j(\lambda)$ given by,

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

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 $\hat{\beta}_j^{NG}(\lambda) = d_j(\lambda)\hat{\beta}_j^{LS}$, $j = 1, \ldots, p$.

Group LASSO

- Extension of LASSO.
- For a vector $\eta \in \mathbf{R}^d$, $d \geq 1$, and a symmetric $d \times d$ positive definite matrix K, we denote that,

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

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 of K_j is to let $K_j = p_j I_{p_j}$, $j = 1, \ldots, J$, such that

$$\lambda \sum_{j=1}^{J} \|\beta_j\|_{K_j} = \lambda \sum_{j=1}^{J} (p_j \sum_{i=1}^{p_j} \beta_{ji}^2)^{1/2}.$$

Group LASSO (Continued)

- The objective function of group LASSO is constructed by a quadratic loss and a group LASSO penalty.
- The group LASSO penalty is 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.
  - At the group level, it is l_1-penalty, penalizing the magnitude of a group (i.e., $(p_j \sum_{i=1}^{p_j} \beta_{ji}^2)^{1/2}$), and the groups with small magnitude will be shrunken to 0, and thus it enables group selection.
  - Within the group, the penalty is essentially the sum of squares of coefficients, so that 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.

Group LASSO (Continued)

Blockwise Gradient Descent Algorithm

- Each iteration uses block soft-thresholding to provide a closed form solution.
  - 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 $X_j' X_j = I_{p_j}$, 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 $j = 1, \ldots, J$.

Group LARS

- Extension of LARS.
- We start by considering a simplified setting of group selection where $p_1 = \ldots = p_J = p$.
- Define the angle $\theta(r, X_j)$ between an n-vector of 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.
- $\cos^2\{\theta(r, X_j)\}$ 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 the group LARS modifies the LARS algorithm only by looking for the vector, say X_j, that has the smallest angle with Y (i.e., $\|X_j' Y\|^2$ the largest).

Group LARS (Continued)

- **Group LARS algorithm in general setting:**
  1. Start from $\beta^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$.
  2. Compute the current 'most correlated set', $\mathcal{A}_1 = argmax_j \|X_j' r^{[k-1]}\|^2 / p_j$.
  3. 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 $\mathcal{A}_k$.
  4. 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 $\mathcal{A}_k$.
  5. 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 $\alpha = 1$.
  6. 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 $\alpha = 1$.

Group Non-negative Garrotte

- Extension of Non-negative Garrotte.
- Under the setting of group selection, $\hat{\beta}_j^{LS}$ is a vector, and every component of $\hat{\beta}_j^{LS}$ is scaled 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.

- **Group Non-negative Garrotte algorithm:**
  1. Start from $d^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$.
  2. Compute the current active set $\mathcal{C}_1 = argmax_j(Z_j' r^{[k-1]}/p_j)$.
  3. Compute the current direction γ, which is a p-dimensional vector defined by $\gamma_{\mathcal{C}_k^c} = 0$ and $\gamma_{\mathcal{C}_k} = (Z_{\mathcal{C}_k}' Z_{\mathcal{C}_k})^- Z_{\mathcal{C}_k}' r^{[k-1]}$.

Group Non-negative Garrotte (Continued)

- **Group Non-negative Garrotte algorithm (Continued):**
  4. 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 $\mathcal{C}_k$.
  5. 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.
  6. 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 $\mathcal{C}_{k+1} = \mathcal{C}_k - \{j^*\}$.
  7. Set $r^{[k]} = Y - Zd^{[k]}$ and $k = k + 1$. Go back to the third bullet until $\alpha = 1$.

Tuning

- A simple approximate C_p-type criterion is introduced to select the final estimates.
- 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,

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

- However, it consists of unknown values...

Tuning (Continued)

- The following approximations are proposed:
  - For group LASSO,

$$\tilde{df} = \sum_j I(\|\beta_j\| > 0) + \sum_j \frac{\|\beta_j\|}{\|\beta_j^{LS}\|}(p_j - 1)$$

  - For group LARS,

$$\tilde{df} = \sum_j I(\|\beta_j^{[k]}\| > 0) + \sum_j \left(\frac{\sum_{l<k} \|\beta_j^{[l+1]} - \beta_j^{[l]}\|}{\sum_{l<J} \|\beta_j^{l+1} - \beta_j^{[l]}\|}\right)(p_j - 1)$$

  - For group non-negative garrotte,

$$\tilde{df} = 2\sum_j I(d_j > 0) + \sum_j d_j(p_j - 2)$$

- 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.

Other Properties

- The group LASSO and group LARS are equivalent when the full design matrix X is orthogonal, but can be different in more general situations.
- The solution path of the group LASSO is generally not piecewise linear whereas the solution path of group LARS is.

Summary

Advantages

- 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.

Summary (Continued)

Limitations

- 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 approaches should be used. An example would be the group bridge as proposed by Huang *et. al.* (2009).
- 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.

References

- 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.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics, 32*(2), 407-499.
- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics, 37*(4), 373-384.
- Huang, J., Ma, S., Xie, H., & Zhang, C. H. (2009). A group bridge approach for variable selection. *Biometrika, 96*(2), 339-355.