Salil Khandelwal
STAT 852
Modern Methods in Applied Statistics
301376050

Q1 95

Project 2

Group LASSO

Introduction

Variable Selection is an important part of modeling a dataset. Not everything that is measured helps in predicting the response variable. Sometimes we are better off selecting a subset of variables. It is a question of bias-variance tradeoff. If we have too many variables it will result in models having high variance and if we have too few variables it will result in the model having a high bias. This is where variable selection techniques come in. Previously, we have seen methods that can be grouped as classical and modern techniques of variable selection. They are:

1. Classical: Stepwise Regression, All-Subset Regressio
2. Modern: LASSO and its variants.

(Of course, there are other methods in each group, but we didn't cover them.)

Methods like stepwise regression tend to be greedy and often reacts strongly to the signal. LASSO is considered the state of the art method for variable selection. Though LASSO does a fairly good job, it does not take into account the inherent grouping present within the variables. A prominent example of this is when we are creating dummy variables for categorical predictors. The variables representing differe are intrinsically “grouped” . If LASSO is run on the set of variables, it may select certain levels of the dummy varia picture of whether that categorical predictor is actually us

Some might argue that if any level is important, then the variable is important. However, the selection of individual dummies may depend on the baseline value chosen. Specifically, if the baseline level has an average mean or small number of obs, then comparisons to other levels are inherently small. If it is chosen to be one with an extreme mean value and/or a lot of cases, then comparisons will seem inherently large.

Consider the following example for more clarity. In Project 1 we have seen that the dataset had fifteen variables (X1-X15). Out of these variables X1, X2 and X15 were categorical and dummy variables were created for them. When variable selection techniques were ran, it was observed that certain levels were being selected, for instance X152 (2nd level of X15), X1512(12th level of X15) etc. and some levels were neglected. Now our dilemma was should we include X15 or not as our important variable. At the time of project we decided that if any level of categorical variable is being selected often, we will select the entire group of that variable(including all levels). Through this procedure we ended up selecting all levels of X2 and X15 along with X4 and X12 as important variable.

However, an elegant way of handling variable selection in such a premise was introduced by Yuan and Lin, when they extended LASSO to consider inherent grouping present between the variables and coined the term Group LASSO. In this report we are going to understand more about this method.

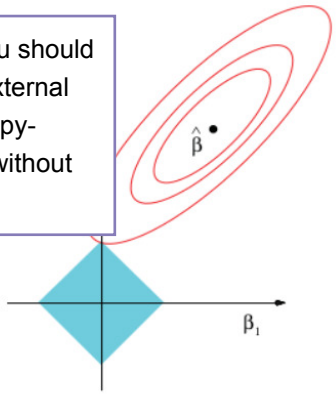
LASSO Review

We are reviewing LASSO as the ideas behind it can help us to transition easily to Group LASSO. The core concept of LASSO is to add a “penalty term” to the OLS criterion.
The parameter estimate for LASSO is:

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

The second part of this equation $\lambda||\beta||_{l1}$ is an L1 norm of the parameter vector and causes sparsity. This can be understood intuitively from the figure below.

If you copy someone else's plot, you should cite them as a source. In fact, for external publication, plots are considered copy-righted and cannot be reproduced without permission of the copyright holder.



Imprecise. The minimum within the constraint region will tend to lie on the axis *as long as the constraint region remains small enough.* When it grows, the optimal parameter estimates (not "variables") eventually all move away from 0.

Consider a simple setting where only 2 variables are there. contours of the error and the blue square depict the constra induced as the minima will lie along the axes causing some LASSO as well we will be doing something similar, we will be coming up with a penalty term and also observe intuitively how sparsity happens at the group level in that method.

Group LASSO

Before explaining the method, we would like to introduce some notation and definitions that will be used throughout this report.

The extended regression equation:

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

- We have J groups of variables.
- Each group j has pj number of variables.
- Xj is a matrix of size (n x pj), where n is the number of observations in the dataset.

(epsilon?)

- βj is the coefficient vector of size pj.

We consider this equation as an extension of the original regression equation in the sense that if all groups are of size 1 it will result in the original regression equation. The response variable and each input variable are centered and the observed mean is 0. Each of the Xj is orthonormal as well to make things simple.

Is this an added assumption, or do we get it for free (without loss of generality)?

The penalty term which can help induce sparsity at the group level is

$$\Sigma||\beta_j||_{Kj} = \Sigma(\beta_j' K_j \beta_j)^{1/2}$$

What is the intuition here?

Here β_j is a coefficient vector of a group containing p_j predictor variables. So $\beta_j \in R^p$ and K_j is a p_j x p_j symmetric positive definite matrix.

The Group LASSO estimate can be obtained by minimizing the following objective criterion.

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

Here λ >=0 and is a tuning parameter much like we have seen in LASSO. $K_1 ... K_J$ are positive definite matrices. A question that comes to mind is what should be the values of K_j. Yuan and Lin have suggested to use $K_j = p_j I_{pj}$. If we plug in the value in the objective criterion above we get:

Why?

$$\min_{\beta \in R^p} \left(||y - \sum_{\ell=1}^{L} X_\ell \beta_\ell||_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} ||\beta_\ell||_2 \right)$$

This is the equation which is in the HTF book and now we know how it is derived. Here L is the total number of groups. If we observe this equation, we can realize that it is a generalization of LASSO criterion we have seen before. To confirm this assume that each

You used J for this before. You should define notation once and convert different notation into yours.

group is of size 1 and the criterion mentioned above becomes equivalent to LASSO criterion.

This acts like LASSO at the group level. Depending on the values of λ, the entire group of predictors can be left out. To understand more intuitively we can look at the geometry of the penalty function much like we did while doing the LASSO. For sake of simplicity consider we have three variables X_{11}, X_{12} and X_2 .The variables X_{11} and X_{12} belong to one group X_1. The coefficient vector of X_1 is $\beta_1 = (\beta_{11}, \beta_{12})'$ and the coefficient for X_2 is simply scalar β_2. Also assume K_j to be I_{pj} (Identity). The expression for the penalty term will be:

$$\Sigma||\beta_j||_{Kj} = \Sigma(\beta_j' K_j \beta_j)^{1/2}$$

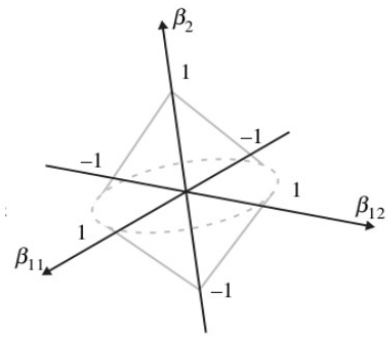
$$= (\beta_1' K_1 \beta_1)^{1/2} + (\beta_2' K_2 \beta_2)^{1/2}$$

$$= (\beta_1' \beta_1)^{1/2} + (\beta_2' \beta_2)^{1/2}$$

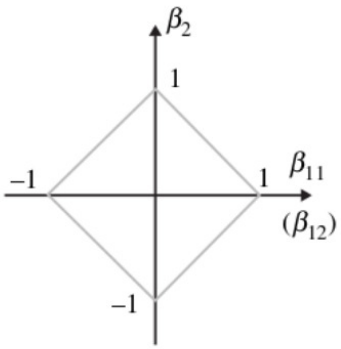
$$= (\beta_{11}^2 + \beta_{12}^2)^{1/2} + \beta_2$$

$$= \beta_{group1} + \beta_{group2}$$ (Denote $(\beta_{11}^2 + \beta_{12}^2)^{1/2}$ as β_{group1} and β_2 as β_{group2})

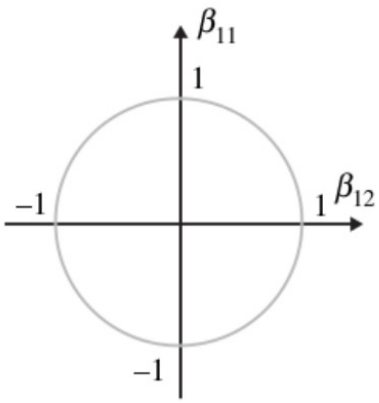
This seems to be the LASSO penalty on group level. Another interesting thing to note is that within the group it seems like a ridge penalty, for example we have $\beta_{11}^2 + \beta_{12}^2$ for group 1 . The Group LASSO penalty can be thought of as acting like LASSO on the group level and like ridge within a group. This thing can be further understood using the figures below



Group LASSO penalty



Intersection of contour with the plane $\beta_{12} = 0$. This looks exactly like the LASSO constraint region and we know this will induce sparsity as the estimate value will lie along the axis. Thus we can see how sparsity is induced at the group level (note this plot is between β_{11} and β_2 which belong to different group)



Intersection of contour with the plane $\beta_2 = 0$. This looks exactly like the Ridge constraint region and we know this will cause shrinkage within a group (note this plot is between β_{11} and β_{12} which belong to same group). Now, we know how Group LASSO causes sparsity at the group level.

We have seen that sparsity is induced at group level, what if we want sparsity within a group as well. This can't be achieved by the Group LASSO criterion we used. In order to get sparsity at both group and individual feature level we have Sparse Group LASSO whose criterion is:

$$\min_{\beta \in \mathbb{R}^p} \left(||\mathbf{y} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \beta_\ell||_2^2 + \lambda_1 \sum_{\ell=1}^{L} ||\beta_\ell||_2 + \lambda_2 ||$$

How? How does it work? Does it select groups and then individuals within groups, or individuals only or...? How does it behave differently from the ordinary LASSO?

It can be thought of as a mix between LASSO and Group LASSO.

R packages and Simulation on Abalone Data

There are three packages that can perform Group LASSO. They are:

1. grpreg
2. gglasso
3. grplasso

In which version? You have presented two. If there is only one lambda, is this fitting the "whole-group" LASSO?

The grpreg package comes with grpreg() and cv.grpreg(). It resembles a lot and cv.glmnet() which we have seen previously and is my preferred method three. cv.grpreg() performs cross-validation for finding out the optimal λ.

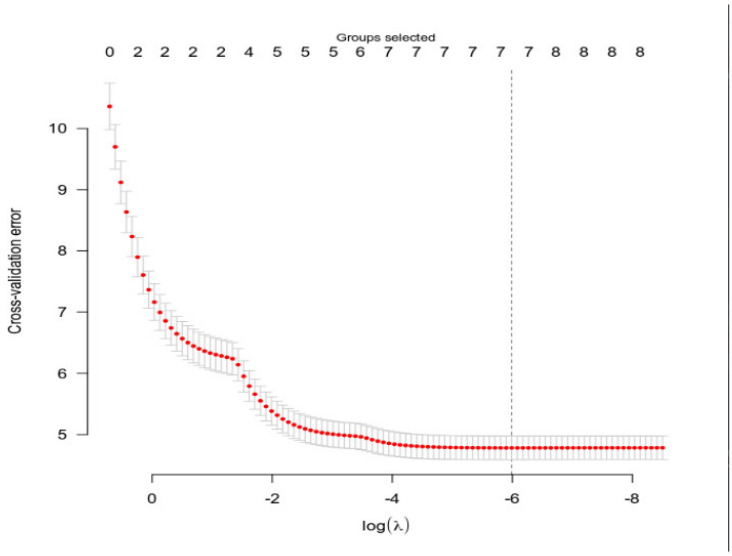
The code snippet showing how to use Group LASSO on the Abalone data:

```
groups = as.factor(c("Sex","Sex","Length","Diameter","Height","Whole","Shucked","Viscera","Shell"))

cv.fit <- cv.grpreg(X=train_X, y=train_y, group=groups, penalty = "grLasso", family="gaussian")
```

"group" argument in the cv.grpreg() and grpreg() is used to define the group of variables. For instance we want first two variables (Male & Female) in the same group Sex we can define it as shown above.

So we must first define the indicators. Does this matter?



A plot showing process of cross-validation and indicating that 7 groups are selected out of 8 possible groups(Length was excluded).

We ran 20 splits on the Abalone data and compared the performance of Group LASSO with other methods seen previously.

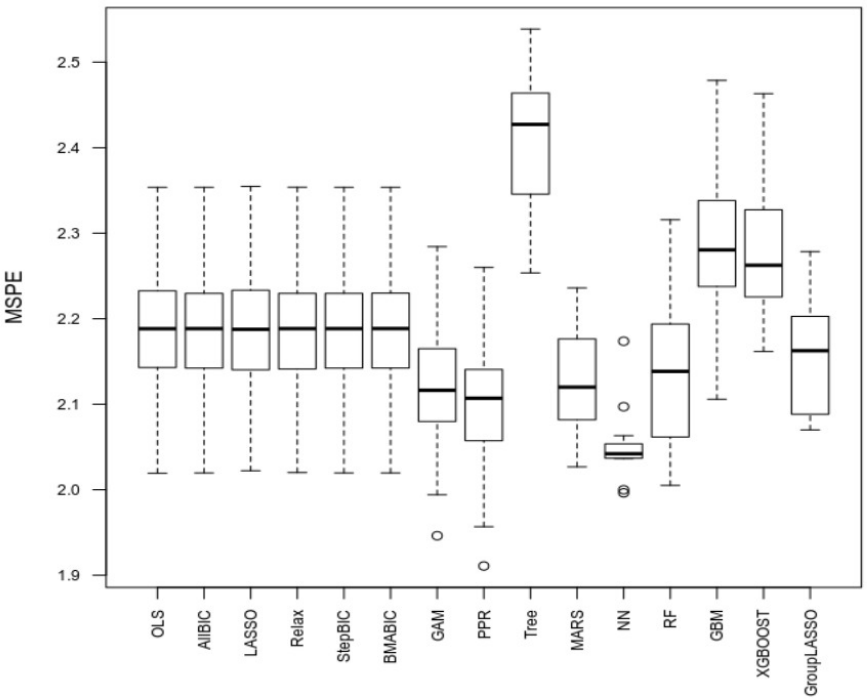
The most interesting comparison would have been against the regular LASSO. There is only so much that a linear regression method can do here, but does the added opportunity to group sexes help us?

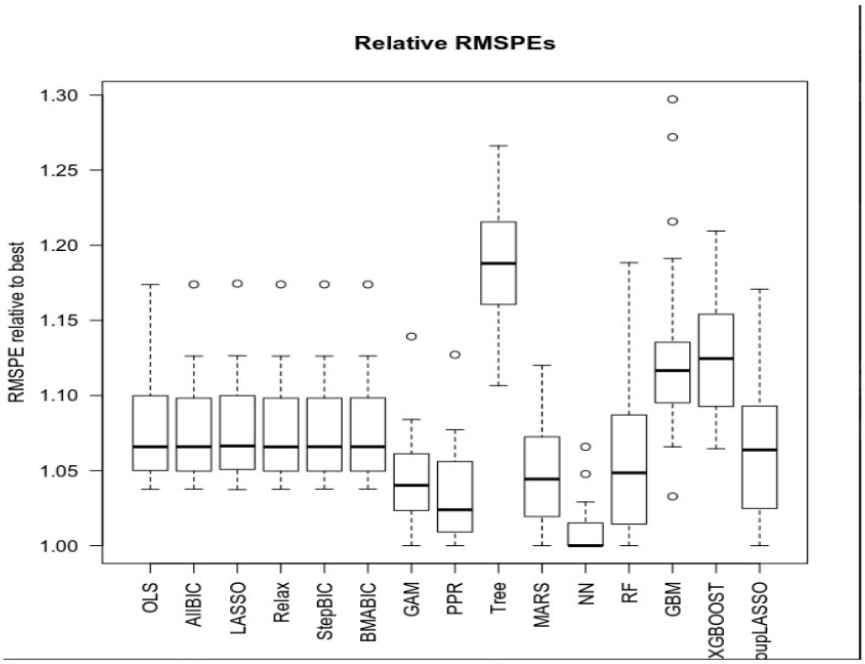
The plots for MSPE and Relative MSPE are given below. W
LASSO is doing a decent job for this dataset(look at the ex
Network is still the best performer on the dataset. Group LA
and xgboost and is giving somewhat similar performance t

A question was raised during the presentation regarding th
variables and how does it change the result when we set d
experimentation on this dataset we can also conclude that setting different baseline gives the same results for Group LASSO.

The way to check this would be to run a non-randomized (non-CV) version of LASSO and compare results. Are they identical? If you reran the Abalone comparisons with CV selection of lambda and got "comparable" results (but not identical due to randomness in CV), then this leaves the answer less clear..

BTW, the paper does say that the choice of orthonormalization should not matter. So we should be able to input any version of contrasts that cover the same column space of the groups.


MSPE
2.5
2.4
2.3
2.2
2.1
2.0
1.9
OLS
AllBIC
LASSO
Relax
StepBIC
BMABIC
GAM
PPR
Tree
MARS
NN
RF
GBM
XGBOOST
GroupLASSO


Relative RMSPEs
RMSPE relative to best
1.30
1.25
1.20
1.15
1.10
1.05
1.00
OLS
AllBIC
LASSO
Relax
StepBIC
BMABIC
GAM
PPR
Tree
MARS
NN
RF
GBM
XGBOOST

Summary

In this report, we learned about Group LASSO and how it induces sparsity at the group level. We gained an intuitive understanding of why it works. We also saw how Group LASSO can be considered as a generalization of the LASSO approach. It is a good tool to have especially when we need to perform variable selection where there is inherent grouping present within the variables. We also very briefly touched upon Sparse Group LASSO. The simulation was shown on the Abalone data and performance was compared with previous methods.

References:

1. STATISTICS 852 – Modern Methods in Applied Statistics, The LASSO: A Modern Approach to Variable Selection - Thomas M. Loughin

2. Model selection and es Ming Yuan, Yi Lin

3. The Elements of Statis oshirani, and Trevor Hastie

4. A note on the group las man, Robert Tibshirani, and Trevor Ha

Source of figures:

1. The contour images ar

Report: 83/100:

- Relation to classwork mostly clear, but a few minor inaccuracies. 9/10
- Overall comprehension of the topic is OK. This was a surface treatment, which is OK, but the descriptions stayed pretty close to published sources. The intuition behind ideas was not given as much attention as I would have liked. 46/50
- Accurate reporting. 30/30
- Appropriate use of the class example, with a few minor concerns. 9/10

Presentation: 94/100 Good overall talk. Good choice of material to present and good slides. Just OK with presenting ideas and intuition, and just OK with explanations for questions.

Difficulty: +1 This is a harder paper if covered in detail, but a surface treatment is more moderate.

OVERALL 94(0.75) + 93(0.25) +1 = 95 +0.25 participation