Group lasso: Wrong vs correct

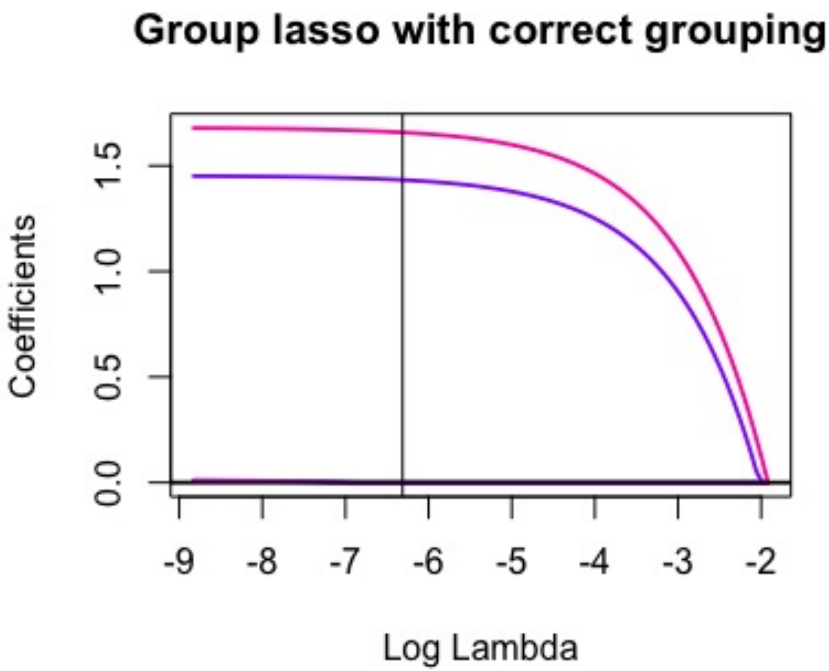
May 2019

Simulating Data

- Data is simulated from the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of responses, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the matrix of features, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of coefficients and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a vector of residual noise.
- Simulating data with n $= 500$ observations and p $= 20$ variables.
- $\boldsymbol{\beta}$ is simulated from a $N(0, \mathbb{I}_p)$- distribution, and every entry in $\boldsymbol{X}$ from a normal distribution with $\mu = 0$ and $\sigma = 1/\sqrt{p}$.
- Dividing $\boldsymbol{\beta}$ into four groups groups (1,2,3,4) with 5 coefficients in each group according to the following grouping scheme: $\{(1,1,1,1,1),(2,2,2,2,2),(3,3,3,3,3),(4,4,4,4,4)\}$ and then setting all coefficients belonging to group 1 and 3 to zero.
- Noise $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbb{I}_n)$ is simulated with target SNR $= 20$.

Group lasso with correct groups.

- Using the gglasso and cv.gglasso functions in R, giving the functions the correct groups to see how it performs.
- Using cross validation with k =5 to find the optimal penalty lambda.



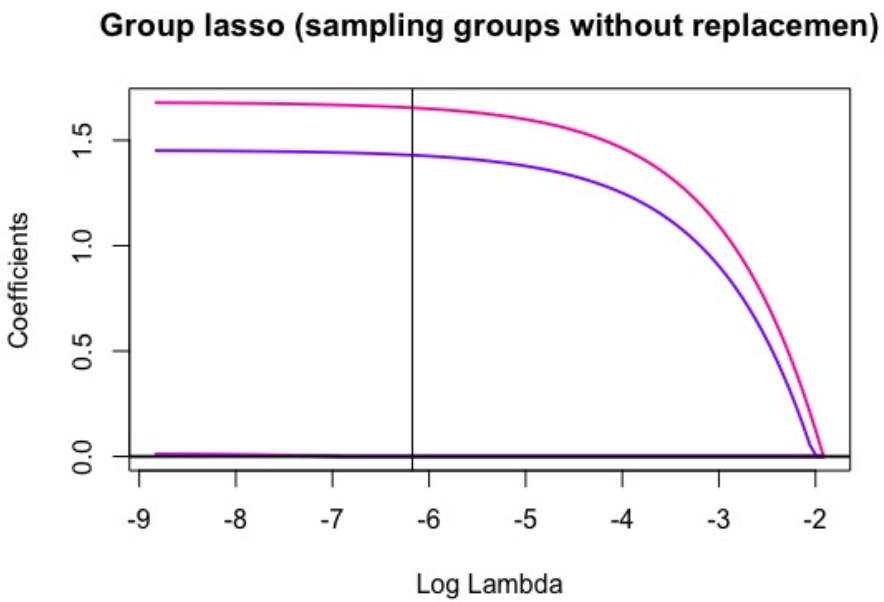
- Straight black line is the "optimal" lambda.
- As we can see, the lasso procedure correctly sets groups 1 and 3 to zero for the optimal lambda.

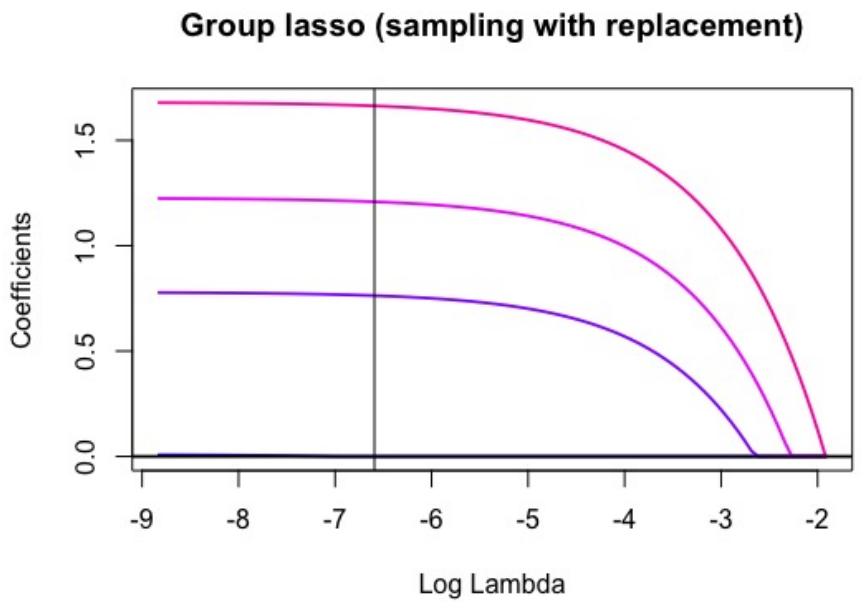
Group lasso with incorrect groups

- What happens when we pass the wrong group assignments to the lasso procedure?
- Two different procedures of creating "wrong" grouping assignments. First randomly sampling 20 values WITHOUT replacement from the original grouping scheme, basically just scrambling the grouping.
- Second procedure we instead randomly sample WITH replacement. So different groups can now have different number of coefficients.
- The procedures gives us the two different groupings:
- without replacement: $\{1, 3, 2, 1, 2, 3, 4, 3, 1, 1, 2, 4, 3, 2, 4, 2, 4, 4, 1, 3\}$
- with replacement $\{4, 4, 2, 2, 2, 2, 3, 2, 2, 4, 2, 2, 1, 3, 3, 1, 2, 2, 4, 3\}$

Group lasso with incorrect groups

- As we can see in the left figure, the method without replacement manages to correctly set group 1 and 3 to zero for the optimal lambda.
- When using the method without replacement, the gglasso procedure only sets group 1 to zero for the optimal lambda.



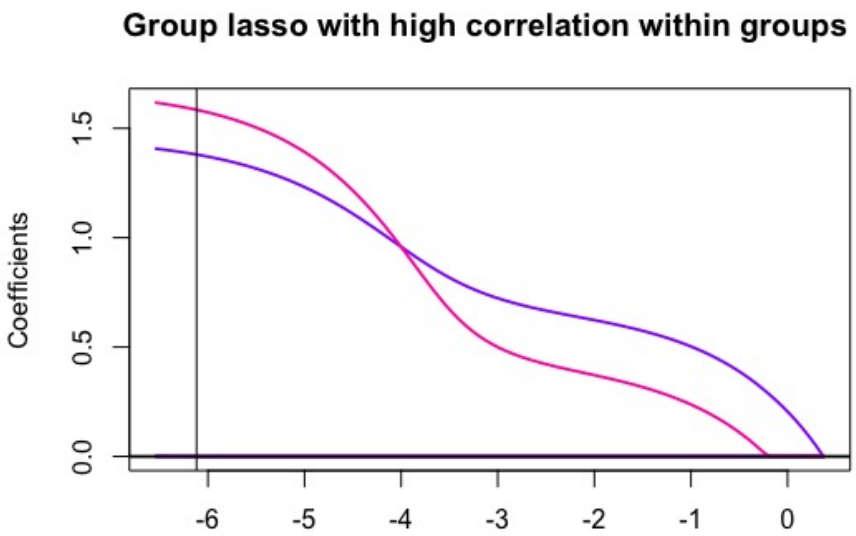


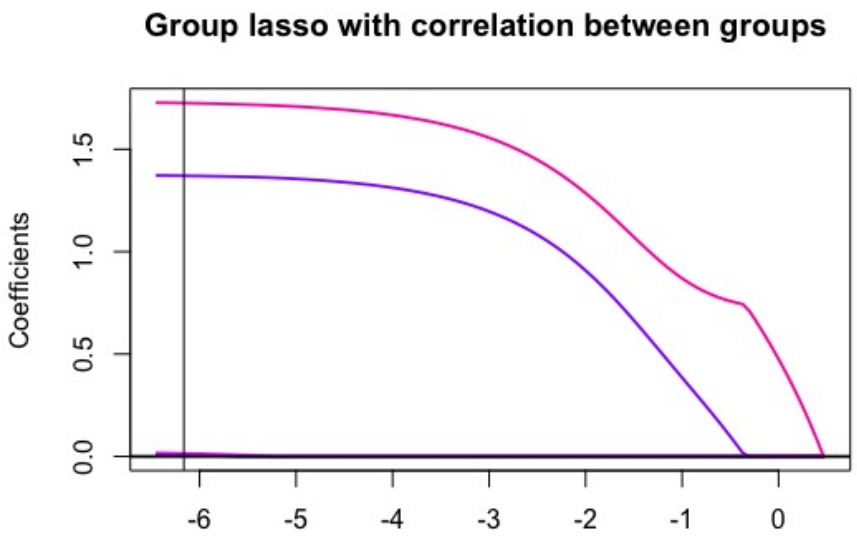
Group lasso with incorrect groups

- The group lasso seems more robust to a false grouping if the original groups have been "scrambled" or systematically reorded.
- Even though we give it the wrong grouping, it still manages to correctly set the prespecified groups that were zero to zero and vice versa.
- However, when we randomly group with replacement. That is, when we change the whole structure of the grouping by letting more (or less) coefficients belong to a group. The lasso procedure is not as robust.

Exploring the effect of correlation within and between groups

- The left plot shows the results of the group lasso when simulating data with a correlation of 0.95 within groups and 0 correlation between groups.
- Right plot instead shows when simulating data with a correlation of 0.5 within groups and 0.6 between groups.
- Both procedures have been given the correct groups from the first slide.





Exploring the effect of correlation within and between groups

- High correlation within groups, makes the group lasso procedure effective. This makes sense since we have obviously (in a real setting) grouped variables by some sort of similarity or intuition based on the data. We want certain variables to be either in the model or not, and if the variables in a group are highly correlated it is easier for the group lasso procedure to separate and distinctify the different groups.
- When including correlation within groups and between groups we get a different result. As we can see (maybe hard since 2 of the groups are pretty close to zero) in the right plot on the previous slide, for the optimal lambda actually none of the groups are set to zero. It seems like the group lasso procedure gets a bit "confused" when we have correlation within and between groups and it is not as effective when given the correct groups.