Variations on the LASSO
Aimed at Incorporating Covariate Characteristics

Daniel Grant
(ID: XXXXXXXX)

XXXX XXX Written Report
April 22, 2021

Proposed by Tibshirani (1996), the LASSO is a tool that can be used to find and quantify the effects of important predictors from an often much larger starting set of predictors. In essence, this method is based on finding coefficients subject to the minimization problem

$$\underset{\beta}{argmin}\frac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1$$

In this method, no information about the relationship between covariates is utilized, using only the notion of minimizing the residual sum of squares of a linear model, along with an ℓ_1 penalty for the coefficients β[5]. While this method may work well at many of the tasks it is used for, there are situations where using additional predictor information can result in better performance in terms of prediction error and in terms of selecting domain-relevant variables, when compared to traditional LASSO. There are 4 such main areas where covariate characteristics may be used to improve upon the traditional LASSO. The following review will discuss these suggested improvements, and additionally discuss their comparison to the traditional LASSO.

Grouped Observations

The first such case involves samples in which predictors come from non-identical subgroups of data. For example, there may exist patients with differing classifications of the same disease. If the goal of an analysis is to determine which predictor variables are influential and to what degree while simultaneously accepting that these variables may not have identical effects in each subgroup, the traditional LASSO is not formulated to handle this problem; coefficient estimates and selected variables are based on the entire sample that the method is applied to and do not make a distinction between subgroups of the data (without performing LASSO on each subgroup individually). Dondelinger and Mukherjee (2018) provide a solution to this problem via their joint LASSO, with estimates subject to

$$\underset{[\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_K]}{argmin}\sum_{k=1}^{K}\left(\frac{1}{n_k}||y_k - \mathbf{X}_k\boldsymbol{\beta}_k||_2^2 + \lambda||\boldsymbol{\beta}_k||_1 + \gamma\sum_{k'>k}\tau_{k,k'}||\boldsymbol{\beta}_k - \boldsymbol{\beta}_{k'}||_2^2\right)$$

where there are K subgroups. In this case, a fusion penalty is amended to the traditional LASSO minimization problem (for which squared residuals are now weighted by sample size per group) which takes into account the proposed similarity between each pair of groups [p.220][1]. Dondelinger & Mukherjee show that this method can in some situations improve on the alternatives of either a LASSO approach ignoring subgroups, or creating individual LASSO models for each subgroup [p.228-233][1]. Since the traditional LASSO does not take this subgroup structure into

account, it is easy to see why taking these subgroups into consideration can be helpful if they do, in fact, exist. While this method also allows for the ability to see whether certain predictors are important in all or only some subgroups, Dondelinger & Mukherjee note that this solution is often much less sparse than either individual or a single aggregate LASSO[p.229-30][1]. In this way, despite the claims of decreased prediction error, it is clear that this method could result in its own interpretability challenges if many predictors are selected.

Grouped Variables

In some situations it is not the groups of subjects that are of interest, but rather proposed groups of covariates within the subjects. This type of consideration may be made particularly often in cases of genetics research, in which predictors may group together to form a larger structure like a gene or pathway [p.49][8]. The need for this approach may also arise as a result of dummy-coding a categorical variable with many levels. In a traditional LASSO, such a categorical predictor may be important (having important levels), but have many of its dummy coded levels removed due to LASSO's variable selection[p.49][8]. In this case, the group LASSO as proposed by Yuan & Lin (2006) has been proposed with covariates selected by

$$\underset{\beta}{argmin}\frac{1}{2}||y-\sum_{l=1}^{m}X^{(l)}\beta^{(l)}||_2^2+\lambda\sum_{l=1}^{m}\sqrt{p_l}||\beta^{(l)}||_2$$

Where p_l is the size of group l and $\beta^{(l)}$ represents the coefficients of predictors in group l. In this way, the penalty term penalizes groups together, ensuring that predictors in each group are treated similarly with respect to variable selection. For a goal of variable selection in data sets with factors, Yuan & Lin found that their method performs similarly to other similar non-LASSO group-based methods (which all performed much better than least squares with backwards stepwise variable selection)[8].

Abundance of Covariates

One of the most common goals of using the LASSO is to reduce the number of selected predictors when the number of initial predictors p is much larger than the sample size n. As Zou and Hastie (2005) note, however, it is only possible to select up to n variables [p.302][9]. They go on to explain that in many contexts, like that of genetics research, the number of genes is very large relative to the sample size, with more than n genes potentially being important to an outcome of interest. Wang et al (2011) propose a solution to this problem via the random LASSO. In this

method, bootstrap samples are taken of the original data, and for each sample, the traditional LASSO is applied with non-zero coefficients being the intersection of those that the method selects and those not in a random previously selected subset of predictors. An importance measure for each predictor is then constructed based on these estimates, which defines the inclusion probability in a second bootstrap sample that includes only some of the predictors. The LASSO is then performed on the second bootstrap sample, and the final effect estimates are averaged over this bootstrap sample. Because of bootstrap sampling, the number of selected predictor variables is no longer bounded by the sample size[7].

In a similar way, there may also be situations in which most predictors are not important. While the group LASSO discussed previously is able to use grouped data to select groups of predictors, Simon et al. (2013) note that there may be situations where only a subset of predictors in selected groups are important [p.49][4]. This lead Simon et al. to develop the sparse group LASSO. This method expands upon the group LASSO by combining both the group LASSO penalty and the traditional LASSO penalty,

$$\underset{\beta}{argmin}\frac{1}{2}||y-\sum_{l=1}^{m}X^{(l)}\beta^{(l)}||_2^2+(1+\alpha)\lambda\sum_{l=1}^{m}\sqrt{p_l}||\beta^{(l)}||_2+\alpha\lambda||\beta||_1$$

Where α is fixed and λ is to be tuned. Simon et al. found suggest that this method has the ability to outperform the traditional LASSO, especially in cases with many groups. While they note that this method can have computation times that are relatively high for a collection of λ values, Ida et al. (2019) propose an alternate optimization algorithm that they claim preserves the accuracy of that of the original while also decreasing computation time significantly [3].

Correlated Variables

One of the most important features of predictor variables are their potential correlations. Since one of the goals of the LASSO is to find a subset of a (often much larger) group of predictors that predicts an outcome, it seems sensible that in many cases predictors may be correlated in some way. As Zou & Hastie (2005) note, correlations between predictors especially in genetics data can often be quite high, and this can often times result in only a single selected variable from a group of correlated variables [p.302][9]. This is problematic if the goal of an analysis is to discover which variables relate to the outcome, if potentially scientifically relevant predictors can eliminated in favor of other irrelevant predictors simply because they are correlated. Because of the frequency of this problem, there have been several variants on the LASSO that attempt to

tackle it. In addition to its ability to select a larger number of predictors than the size of the original sample, the random LASSO by Wang et al. is one such method[7]. This is because the random selection of predictors results in correlated predictors being split up in some bootstrap iterations, allowing for effect estimates for one variable without being influenced by another correlated variable [7].

Another method to tackle this problem directly through the use of a novel regularization term rather than a carefully constructed application of the traditional LASSO is that of the precision LASSO by Wang et al.(2019). Their goal in proposing their new regularization penalty was to select from correlated predictors in a stable way, while simultaneously attempt to solve problems that the traditional LASSO has when predictors are linearly dependent. Their minimization problem takes ideas from the trace LASSO by Grave et al (2011) to solve these two problems simultaneously,

$$\underset{\beta}{argmin}\frac{1}{2}||y - X\beta||_2^2 + \lambda \left|\left|\left[\gamma(X^TX)^{1/2} + (1-\gamma)(X^TX + \mu I)^{-1/2}\right]\text{Diag}(\beta)\right|\right|_*$$

where $||\ ||_*$ is the "trace norm" as introduced by Grave et al.[2], and γ is a parameter that represents the focus towards issues of correlation versus linear dependence.

In this case, the term $(X^TX)^{1/2}\text{Diag}(\beta)$ has the goal of taking into account correlations between variables in regularization. In simulation studies, Wang et al. found that this method tends to perform better than the LASSO and slightly better than other alternative techniques in terms of prediction error when correlations between predictors are high, as intended. When correlations between variables are low, however, they found that this model tended to perform worse than the traditional LASSO. Not only this, but they also found that their model was able to select known causally-linked predictors to the outcome compared to other methods including the traditional LASSO. Because of this, they claim that their method is also superior in terms of variable selection relative to other methods[p.1186][6]. Since the traditional LASSO is known to select highly correlated predictors in an unstable way, this method's attempts to limit this should at minimum be admired for its contributions in providing consistent results to model-builders.

References

[1] Frank Dondelinger, Sach Mukherjee, and The Alzheimer's Disease Neuroimaging Initiative. The joint lasso: high-dimensional regression for group structured data. *Biostatistics*, 21(2):219–235, 09 2018.

[2] Edouard Grave, Guillaume Obozinski, and Francis Bach. Trace lasso: a trace norm regularization for correlated designs. *Advances in Neural Information Processing Systems*, 09 2011.

[3] Yasutoshi Ida, Y. Fujiwara, and H. Kashima. Fast sparse group lasso. In *NeurIPS*, 2019.

[4] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

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

[6] Haohan Wang, Benjamin Lengerich, Bryon Aragam, and Eric Xing. Precision lasso: Accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics*, 35, 09 2018.

[7] Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *The annals of applied statistics*, 5:468–485, 03 2011.

[8] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 02 2006.

[9] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005.