Giang Vu

STAT 2270 – Data Mining

Dr. Lucas Mentch

Nov 6, 2020

Project Write-up

The project is a summary of "Influence of single observations on the choice of the penalty parameter in ridge regression" by Kristoffer H. Hellton, Camilla Lingjærde, and Riccardo De Bin (arXiv: 1911.03662, November 2019).

1. Motivation

What prompted the authors to explore this particular idea stems from the fact that changes in dataset can lead to differences in terms of statistical models, specifically the estimates for coefficients (Breiman et al., 1996; Heinze et al., 2018). Single observations in our dataset can be considered "influential points" when their inclusion/exclusion cause substantial changes in our model (Cook, 1979). And when we look at ridge regression, the choice of the tuning parameter λ (lambda) influences the final model we settle on a lot; therefore, it is important to study how each observation in the dataset can affect the choice of λ (Hellton et al., 2019).

The authors propose a cross-validation based procedure to find the optimal λ, assign weight to each single observation, and look at how the choice of optimal λ varies as a function of the weight of a datapoint. The authors' approach allows us to understand the overall impact of an observation, not just limited to the inclusion/exclusion of it.

2. Ridge regression and cross validation to find tuning parameter λ

The authors consider a linear regression model as follows.

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n.$$

$y_i \in \mathbb{R}$ is the n univariate continuous outcomes, and $x_i \in \mathbb{R}^p$ are p-dimensional covariate vectors. $\beta \in \mathbb{R}^p$ is our vector of regression coefficients and $\varepsilon_i \in \mathbb{R}$ are identically and independently distributed noise terms with mean zero. With design matrix $X \in \mathbb{R}^{n \times p}$ and outcome vector $y = [y_1, \ldots, y_n]^T$

$$\hat{y}(\lambda) = X\hat{\beta}(\lambda) = X\left(X^T X + \lambda I_p\right)^{-1} X^T y = H(\lambda) y$$

is the predictions of ridge regression for a fixed tuning parameter λ, where $H(\lambda)$ is the hat matrix of the ridge regression and I_p is the p-dimension identity matrix. The authors consider leave-one-out cross-validation, with regression coefficients for each fold and prediction error of the i^{th} removed observation being

$\hat{\beta}_{[i]}(\lambda) = \left(X_{[i]}^T X_{[i]} + \lambda I_p\right)^{-1} X_{[i]}^T y_{[i]}$, and $e_{[i]}(\lambda) = y_i - x_i^T \beta_{[i]}(\lambda)$, respectively.

Then the Sherman-Morrison-Woodbury formula for matrix inverses is used to explicitly express the leave-one-out cross-validation error as a function of the residual $e_i(\lambda) = y_i - x_i^T \hat{\beta}(\lambda) = y_i - \hat{y}_i(\lambda)$ as follows

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} e_{[i]}(\lambda)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{e_i}{1 - H_{ii}(\lambda)}\right)^2$$

Where $H_{ii}(\lambda) = x_i^T \left(X^T X + \lambda I_p\right)^{-1} x_i$ is the i^{th} diagonal entry of $H(\lambda)$ (Golub et al., 1979).

From there, we can see the optimal value for λ is

$$\hat{\lambda}_{CV} = \underset{\lambda > 0}{\operatorname{argmin}} \{ \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i(\lambda)}{1 - H_{ii}(\lambda)}\right)^2 \}$$

The optimal tuning parameter minimizes a weighted version of the residuals, where the weights are related to the leverage of the observations via $H_{ii}(\lambda)$. This leads us to the next part to study the relationship between the weights and the choice of optimal λ.

3. Influence of single observations on choosing λ

Weighted cross-validation

For standard leave-one-out cross-validation, the continuously weighted criterion is

$$wCV(\lambda) = \sum_{i=1}^{n} w_i e_{[i]}(\lambda)^2$$

with uniform weights $w_i = \frac{1}{n}$ for $i = 1, \dots, n$. The authors propose exploring the influence of one single observation, say the i^{th} data point, by varying the i^{th} weight only, while the remaining observations each has weight of $w_j = \frac{1-w_i}{n-1}$ for $j = 1, \dots, i-1, i+1, \dots, n$ to make sure all the weights sum up to 1. So then they define the normalized weight cross-validation criterion as

$$wCV(\lambda, w_i) = w_i \left(\frac{y_i - \hat{y}_i(\lambda)}{1 - H_{ii}(\lambda)}\right)^2 + \sum_{j \neq i} \frac{1 - w_i}{n-1} \left(\frac{y_j - \hat{y}_j(\lambda)}{1 - H_{jj}(\lambda)}\right)^2$$

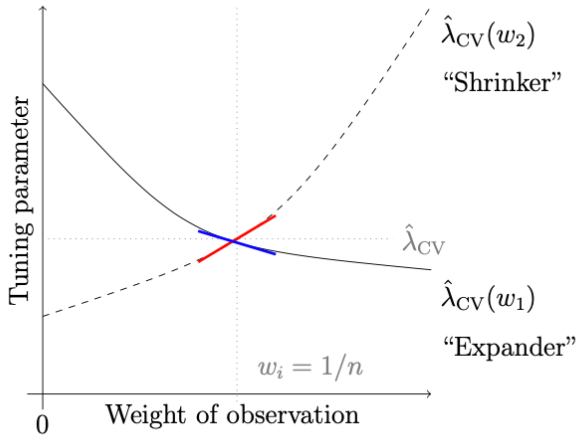
Where w_i can range from 0 to 1, with $w_i = 0$ corresponding to the exclusion of the i^{th} data point and $w_i = 1$ meaning all the other observations are deleted.

Shrinkers and expanders

Using the weighted cross-validation criterion defined above, the authors study how the selection of optimal λ varies as a function of the weight of a single data point, i.e.

$$\hat{\lambda}(w_i) = \underset{\lambda > 0}{\operatorname{argmin}}\, wCV(\lambda, w_i)$$

They then take the derivative of this function with respect to w_i at the specific value $w_i = \frac{1}{n}$ for the case of standard leave-one-out cross-validation, then come up with their classification of two types of observations called “shrinkers” and “expanders”.



Expanders are defined as the observations that make the optimal tuning parameter increase as their weight is decreased; while shrinkers are points that decrease the value of the optimal tuning parameter when their weight is decreased. The visualization of these two classes of points from the paper is included here, where we can see that the weight of expanders has a negative relationship with the value of optimal λ, whereas the weight of shrinkers has a positive with the value of optimal λ (Hellton et al., 2019). Expanders require a smaller optimal tuning parameter, or more degrees of freedom in the model when given more weight; and shrinkers require a higher optimal tuning parameter, or fewer degrees of freedom when given more weight.

By studying the derivative again, the authors also find situations when a point is an expander as follows

- If the i^{th} point has $y_i > 0$, it can only be an expander if $x_i > 0$ (the point lies in the first quadrant) and the OLS coefficient $\hat{\beta}$ must be positive as well (or ridge regression line has a positive slope); or if $x_i < 0$ and $\hat{\beta} < 0$.

- If the i^{th} point has $y_i < 0$, it can only be an expander if $x_i > 0$ and $\hat{\beta} < 0$.; or if $x_i < 0$ and $\hat{\beta} > 0$.

4. Simulation

The authors carry out their simulation study and apply their method to real data, both low-dimensional and high-dimensional cases, and visualize their result by plotting the value of optimal tuning parameter against weight of observation for all the observations in the datasets, along with making a scatter plot of the outcome y_i against the first principal component. Due to limitations of my laptop, I followed the same process with only one low-dimensional dataset, which is a sample from the Boston housing dataset in the R package MASS (Venables & Ripley 2002). My dataset is a random sample of size 100 from 'Boston', which measures property value against various factors of the neighborhood, ranging from crime rate, residential land proportion, to nitrogen oxides concentration and so on. My design matrix contains only continuous factors, all of which are scaled to unit variance. The outcome is median value of houses in thousands of dollars.

Figure 1 below plots the curves of the optimal tuning parameter λ against weights of observations on the left, and on the right a scatterplot of the response against the first principle component (PC1), which explains about 63.2% of the variation for my data.

Points that stand out with the most extreme curves relative to other points are highlighted. Observation 11 (in red) can be identified as a shrinker, because it requires a larger tuning parameter as its weight is increases. We can also see that in the degrees of freedom plot against weights in Figure 2, where the increase of the weight of observation 11 leads to fewer degrees of freedom. Observation 18 (in orange) also stands out as another shrinker, but its

effect on the tuning parameter both when it is down-weighted and up-weighted are not as strong as observation 11.

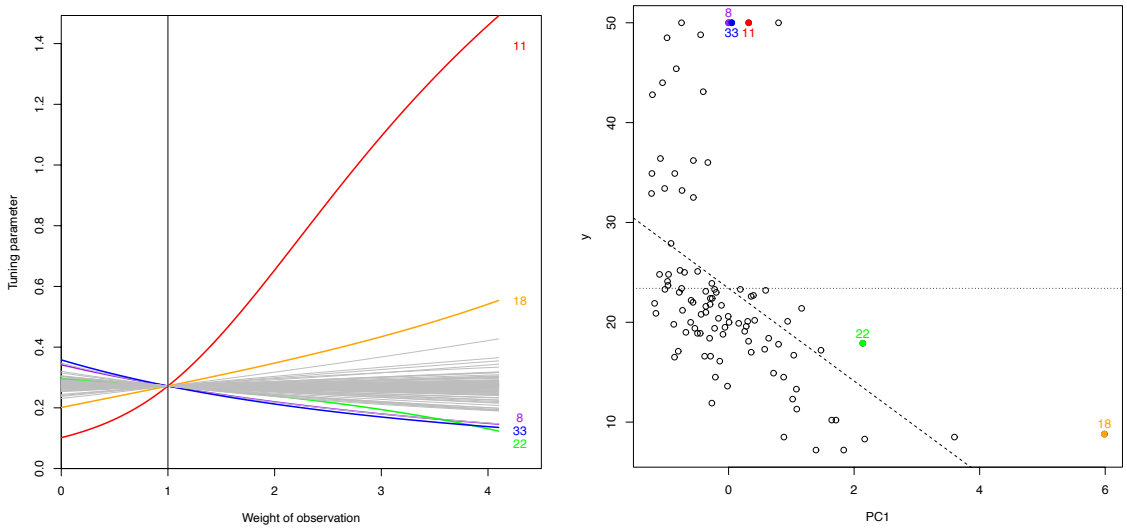


Figure 1: Plot of $\hat{\lambda}_{CV}(w)$ as a function of w_i for the observations in my sample of the Boston dataset (left panel) and scatterplot of the response (median house value) against the first principal component of the design matrix (right panel)

Points 8, 33, and 22 (purple, blue, and green) have expanding effects, as we can see that increasing their individual weight can lead to a decrease in the value of the optimal λ. But none of the expanding effects are strong enough for them to be classified as expanders. We can also see this from the scatterplot to the right of Figure 1, they all lie above the negative-sloped regression line (dashed line) and all have positive x-values and y-values. In Figure 2, we can also see that these points require the most degrees of freedom from the model, i.e requires the most complex models, when their weights are increased.

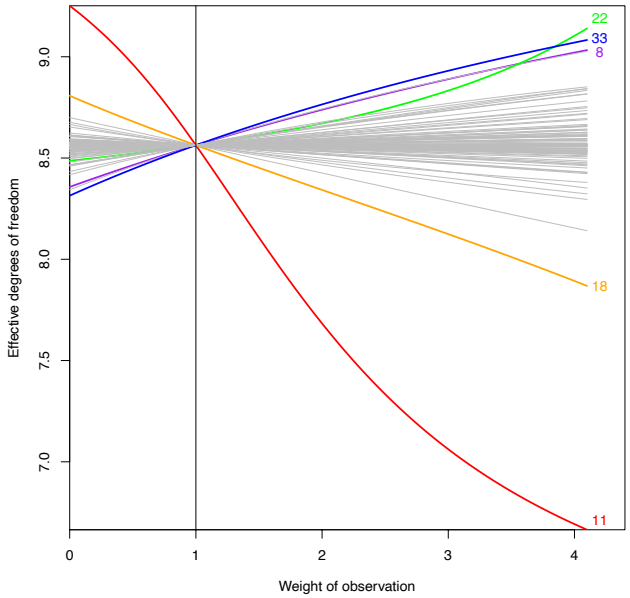


Figure 2: Plot of effective degrees of freedom against the weights of each observation in thee dataset.

One interesting thing to note is that in the authors' simulation, points with expanding effects tend to be very far away from the regression line (i.e. they have large residuals) while points with shrinking effect tend to be close to the regression line. Theoretically this makes sense because a point with a large residual when included in the model will increase model complexity, and the authors also state "the influence of single observations is strongly related to their position with respect to the other points" (Hellton et al., 2019). But in my simulation, the shrinker, observation 11, has a very large residual, almost the same as the residuals of the expanding points 8 and 33, even though observation 11 is supposed to be a point close to the regression line, with similar covariate values (x-values) to 8 and 33, to contrast the effect of both of those expanding points. I repeated my simulation with another sample from the Boston

dataset and still obtained similar results, which do not follow the intuition that the authors propose. I figure there must be some problem with my simulation.

Finally, this method could be used as a tool to identify outliers in the dataset as well. To verify, I calculated the Cook's distance for each data point and the results tell me that observations 11, 18, 22, 30, 40, and 76 are outliers for my dataset. We can see that the authors' method identifies points 11, 18, and 22 as influential points as well. However, this also means that "influential" points in the context of this paper is not the same thing as outliers in the traditional sense. Outliers may or may not be influential in choosing the optimal tuning parameter for ridge regression, and points that have that influence may or may not be outliers in the dataset.

5. Conclusion

Through this project, I get to know about a simple but interesting idea of how single observations in the dataset can have an influence on the choice of the optimal tuning parameter for ridge regression, and how to visualize that overall effect, not just limited to including or deleting a certain point. The method proposed by the authors can also be utilized as a tool for finding outliers in our data. The authors also point out that their intuition behind shrinkers and expanders is still valid for other penalized regression methods, such as lasso, but unfortunately due to time constraint, I have not been able to explore that idea yet.

Reference

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics 24*, 2350–2383.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association 74*, 169–174.

Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics 21*, 215–223.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Kristoffer H. Hellton, Camilla Lingjærde, Riccardo De Bin (2019). Influence of single observations on the choice of the penalty parameter in ridge regression. *arXiv preprint arXiv: 1911.03662*.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0