Project Presentation

STAT 2270 - Data Mining

Dr. Lucas Mentch

Giang Vu - Nov 2020

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

Agenda

1. Motivation
2. Ridge Regression & CV to find tuning parameter lambda
3. Classification of observations' influence on choosing lambda
4. Simulation with real data
5. Conclusion

1. Motivation

- Estimates for statistical models differ when there are changes in the data (Breiman et al., 1996; Heinze et al., 2018).
- One observation in the dataset is considered an influential point when the model difference it causes is substantial (Cook, 1979).
- Because for ridge regression, different values of penalty parameter (lambda) result in very different models, therefore it is worth exploring the influence each single observation in the data set has on the choice of lambda.
- Authors propose using a procedure to find optimal lambda based on cross-validation, assigning weight to single observations, and studying how changes in that weight causes changes in the optimal choice of lambda.
- Authors' approach helps understand the overall influence of a specific data point, not just limited to the inclusion/exclusion of it.

2. Ridge Regression

- Model: $y_i = x_i^T \beta + \varepsilon_i\,, \quad i = 1,\ldots,n\,.$
- Predictions of ridge regression for a fixed tuning parameter λ

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

- Using Sherman-Morrison-Woodbury formula for matrix inverses, authors derive the leave-one-out cross-validation error as a function fo the residual ($e_i(\lambda) = y_i - x_i^T \hat{\beta}(\lambda) = y_i - \hat{y}_i(\lambda)$) as follows

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

2. Ridge Regression

- Note: prediction error of the ith removed observation is $e_{[i]}(\lambda) = y_i - x_i^T \beta_{[i]}(\lambda)$
- Residual of the ith point: $e_i(\lambda) = y_i - x_i^T \hat{\beta}(\lambda) = y_i - \hat{y}_i(\lambda)$
- Optimal value for lambda then will be

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

- It minimizes a weighted version of the residuals (numerator), where the weights are related to the leverage of the observations via the H matrix (denominator).

3. Influence of Single Observations

3.1. Weighted CV

- The authors propose exploring the influence of one single observation, say the ith data point, by varying the ith weight only, while the remaining observations each has weight of $w_j = \frac{1-w_i}{n-1}$ for $j = 1, \dots, i-1, i+1, \dots, n$

- Normalized weight CV criterion is defined as

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

- Authors study the derivatives of this function with respect the ith weight at $w_i = \frac{1}{n}$

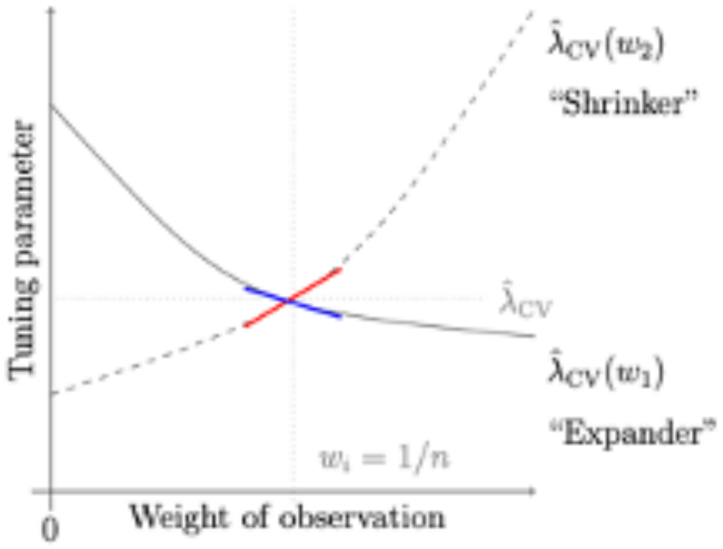
$$\hat{\lambda}(w_i) = \underset{\lambda > 0}{\operatorname{argmin}} \, wCV(\lambda, w_i)$$

3. Influence of Single Observations

3.2. Shrinkers and Expanders

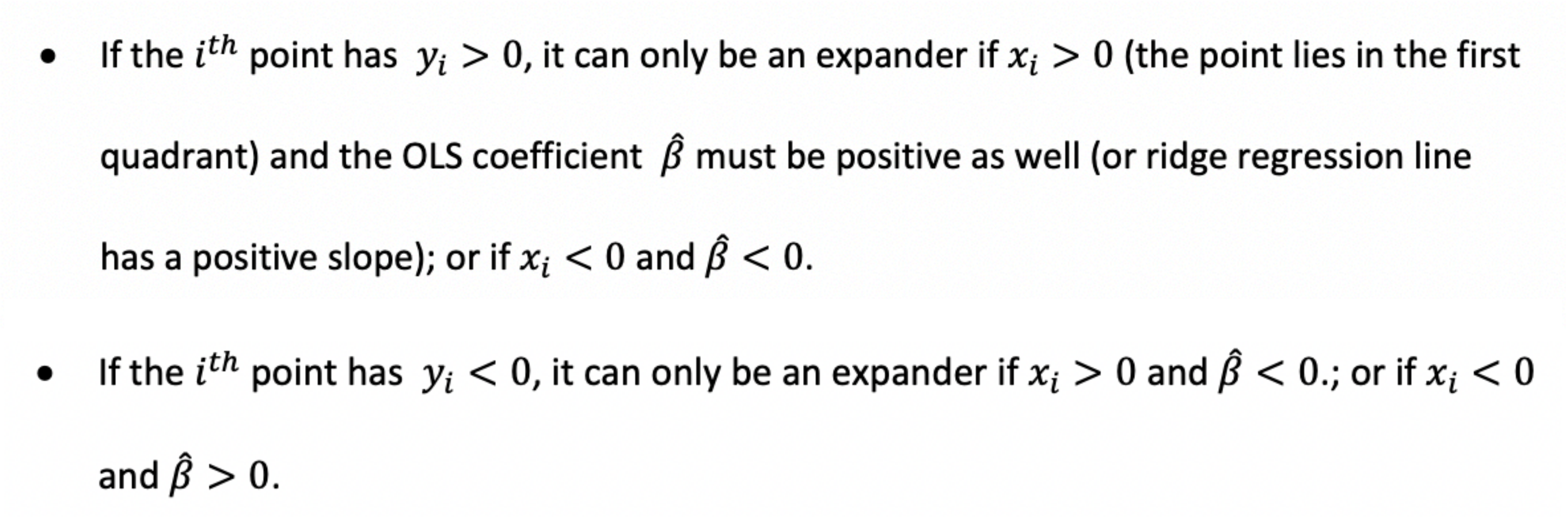
- Expanders are defined as the observations that make the optimal tuning parameter increase as their weight is decreased. Expanders require a smaller optimal tuning parameter, or more degrees of freedom in the model when given more weight.

- Shrinkers are points that decrease the value of the optimal tuning parameter when their weight is decreased. Shrinkers require a higher optimal tuning parameter, or fewer degrees of freedom when given more weight



3. Influence of Single Observations

3.2. Shrinkers and Expanders



- 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

Low-dimensional case

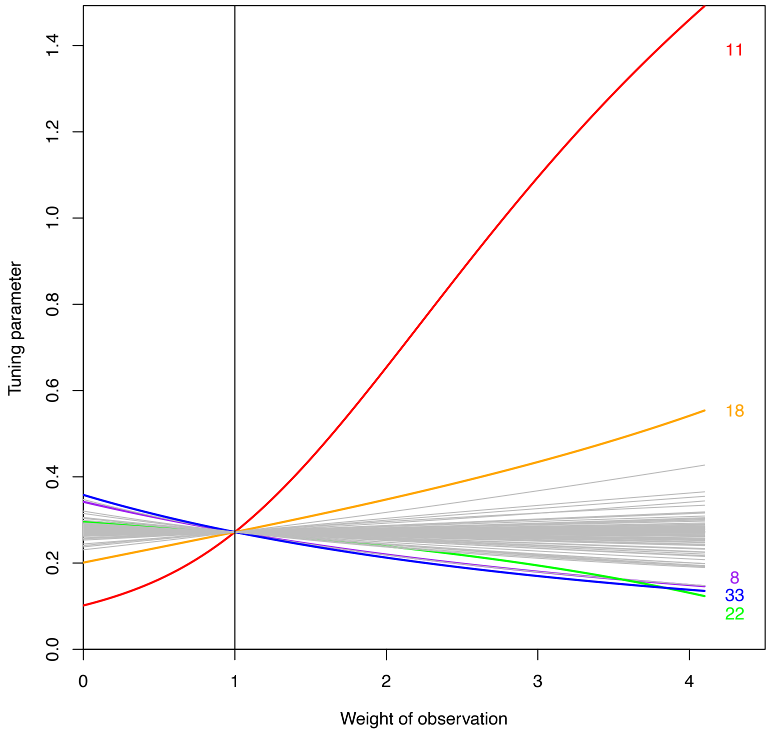
- Data used is a random sample of size 100 from ‘Boston’ dataset from package MASS (Venables & Ripley 2002), which measures property value against various factors of the neighborhood, ranging from crime rate, residential land proportion, to nitrogen oxides concentration and so on.

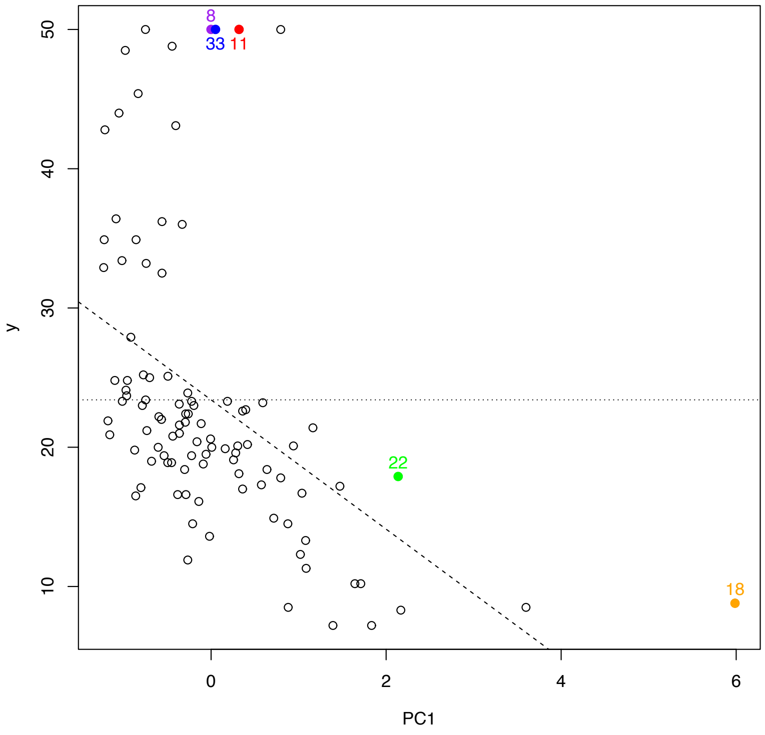
- My design matrix contains only continuous factors, all of which are scaled to unit variance. The outcome is median value of houses in thousands of dollars.

- Using the authors’ process, I plotted the optimal tuning parameter against weights of observations, made a scatterplot of the response against the first principal component, and another plot of weights of observations against effective degrees of freedom required for model.

4. Simulation

Results





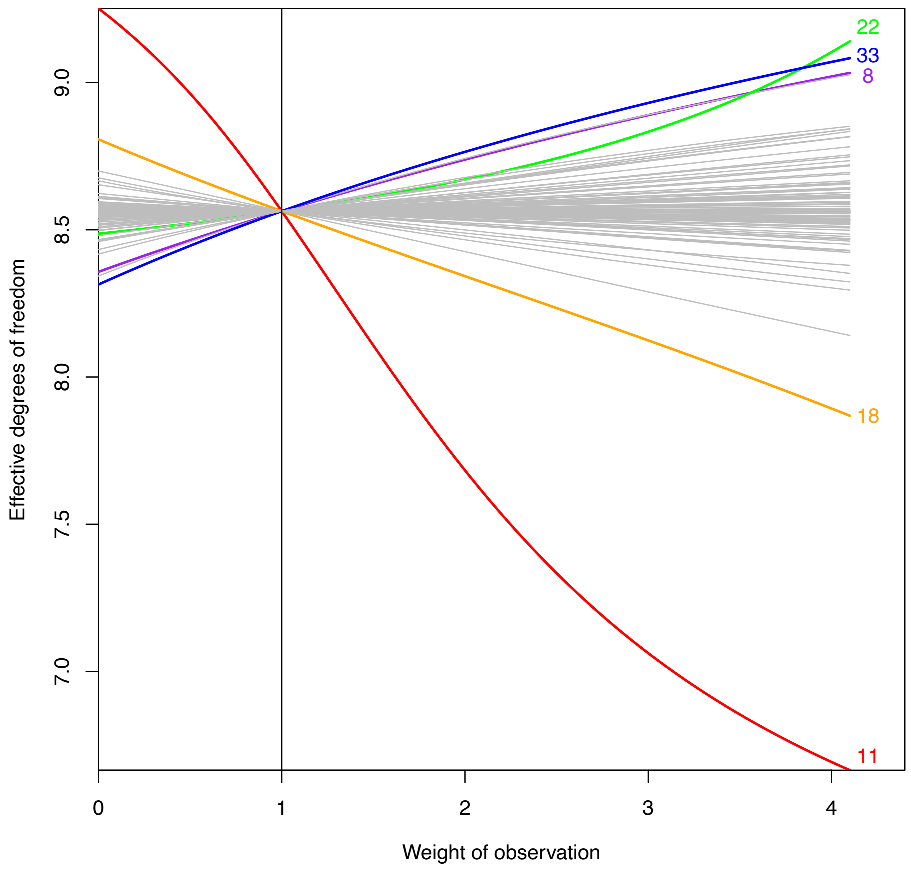
4. Simulation

Results

- Observation 11 (in red) can be identified as a shrinker, because it requires a larger tuning parameter as its weight is increases, i.e. fewer degrees of freedom.

- Observation 18 (in orange) also stands out as a shrinker, but the effect is not as influential as observation 11.

- 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 lambda, they also require more degrees of freedom.



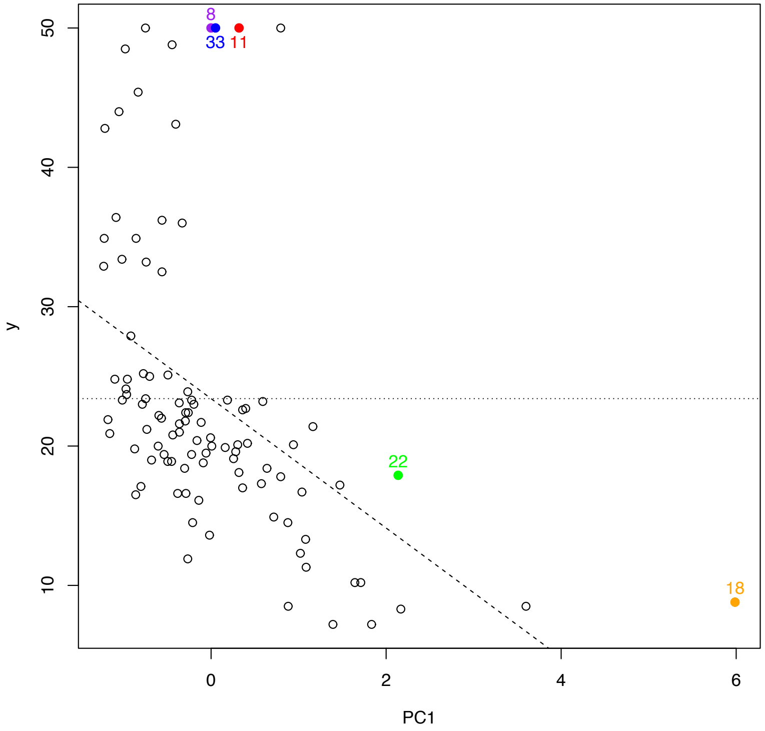
4. Simulation

Notes

- 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.

- 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.

- Redid simulation with different samples but still obtained similar results.



4. Simulation

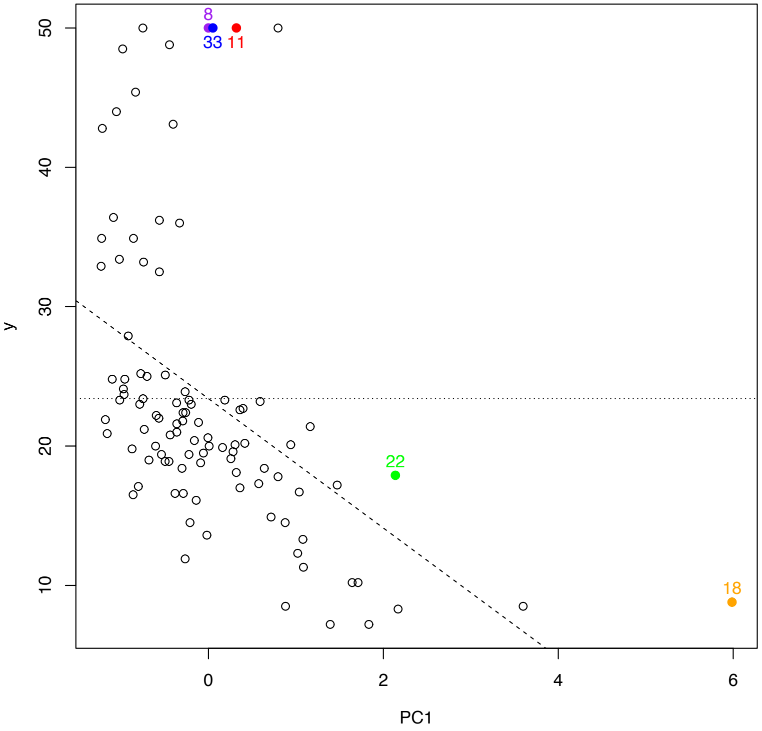
Notes

- Authors' method could be utilized to identify outliers

- Cook's distance gave me points 11, 18, 22, 30, 40, and 76 as outliers.

- The authors' method identifies points 11,18, and 22 as influential points as well.

- 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, and vice versa.



5. Conclusion

- A simple and interesting idea of how single data points can have an influence on optimal tuning parameter's value.
- We can visualize that overall influence as well, not just limited to inclusion/deletion of a certain point.
- Method could be used to identify outliers.
- Authors also state the intuition behind shrinkers and expanders still makes sense in the context of other penalized regression methods, such as lasso, but with non-smooth or non-differentiable tuning parameter curves. However, I haven't been able to explore this idea yet.

References

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. Influence of single observations on the choice of the penalty parameter in ridge regression. *arXiv preprint arXiv: 1911.03662*, 2019.

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