Summary and discussion of: "Effective Degrees of Freedom: A Flawed Metaphor"

Journal club report

LIU, Lingchong

December 15, 2022

1 Summary

In order to demonstrate some irregular properties of degree of freedom, such as non-monotonicity and unboundness, the paper first provided a significant amount of numerical simulation about the concept under the popular definition provided by Efron. A simple example is shown in 1 with the setting

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{X} &= \mathbf{I}_2, \quad \epsilon \sim N(\mathbf{0}, \mathbf{I}_2) \end{aligned} \tag{1}$$

Fitted by the best subset regression.
In this example, the paper showed that even the model is "small" in the sense of $n = p = 2$, the degree of freedom could still be extremely large. With other simulations(will be reproduced in the next section), the paper argued that the common understanding of degree of freedom, i.e. a measure of model flexibility or model size, parametrization of bias-variance trade-off, could be misleading. To clarify the concept of the degree of freedom, and to explain the underlying reasons of appeared irregularities, the paper first convert the model fitting problem to a feasible optimization problem then study the problem from a geometric perspective and finally provide a convincing theoretical explanation.

2 Result and Discussion

2.1 Preliminary

To discuss about the concept of degree of freedom, linear regression model could be a starting point. In a linear regression model, p is the dimension of the feasible set that we are solving β to minimize the loss function. Hence p naturally could be regarded as a measure of the model size and the model flexibility. The generalization of the degree of freedom is motivated by an unbiased estimator of prediction error, proposed by Mallows in 1973[3].

$$E\left(\sum_{i=1}^{n}(y_i^* - \hat{y}_i)^2\right) = E\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + 2\sigma^2 p \tag{2}$$

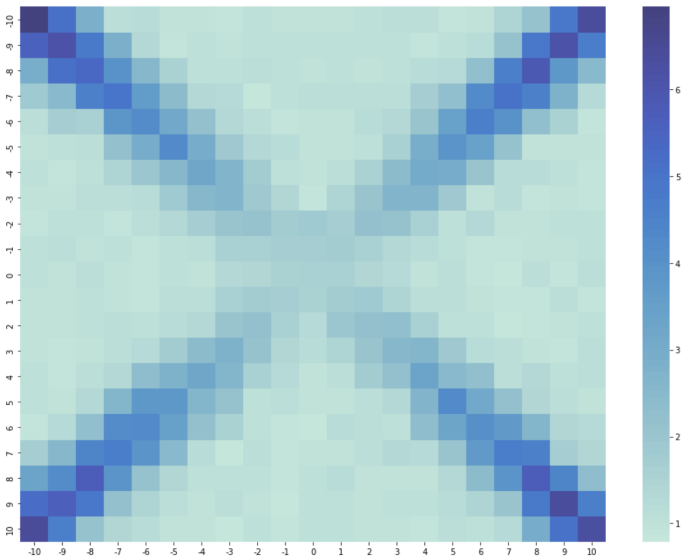

Figure 1: **Degree of Freedom for Best of Two Univariate Models**: The data are generated by a simple Gaussian bi-variate full model with $n = p = 2$. The design matrix X is identity. The axes represent the mean of y and the color in the heat map represents the degree of freedom achieved by a best subset univariate model. (Monte Carlo estimation with 300 simulations)

In (2) y_i^* is an independent random variable sampled at $\mathbf{x}_i$, following the same distribution as y_i. Noticed that in the context of the paper and this report, "variance" of $\mathbf{y}$ **does not contain the variance from sampling**, or from $\mathbf{x}$.
By connecting the formula for general models obtained by Efron in 1986[1],

$$E\left(\sum_{i=1}^{n}(y_i^* - \hat{y}_i)^2\right) = E\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right) + 2\sum_{i=1}^{n} Cov(y_i, \hat{y}_i) \tag{3}$$

The generalized definition of degree of freedom could be

$$DF(\beta, \mathbf{X}, FIT_k) = \frac{1}{\sigma^2}\sum_{i=1}^{n} Cov(y_i, \hat{y}_i) \tag{4}$$

Here FIT_k is the fitted model with hyper-parameter k. β and X are regarded to be fixed.
To estimate the degree of freedom in our simulation, we use Monte Carlo method.

$$\begin{aligned} DF(\beta, \mathbf{X}, FIT_k) &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E\left((y_i - E(y_i))\left(\hat{y}_i - E(\hat{y}_i)\right)\right) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E\left((y_i - E(y_i))\left(\hat{y}_i\right)\right) \\ &\approx \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{1}{J}\sum_{j=1}^{J}\left((y_{i,j} - \bar{y}_i)\left(\hat{y}_{i,j}\right)\right) \end{aligned} \tag{5}$$

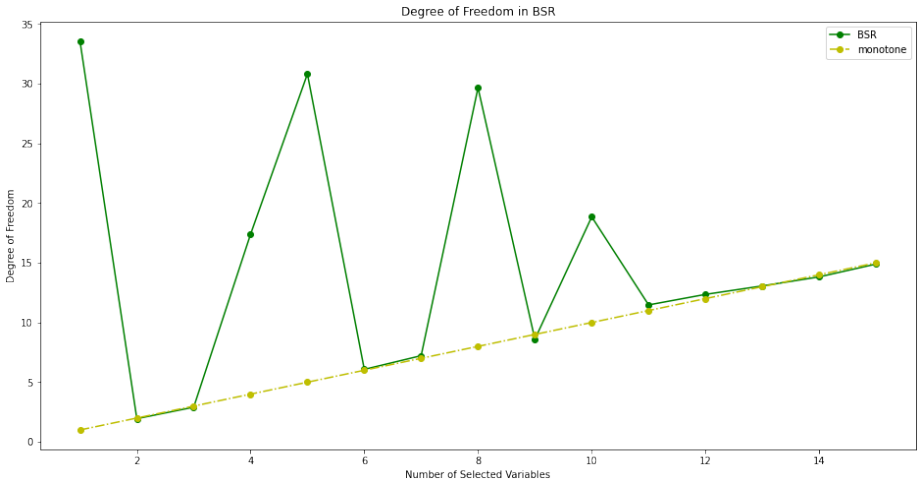

Figure 2: **Degree of Freedom for the Best Subset Models V.S. Subset size**: The data are generated by a linear model with $n = 100, p = 15$. The degree of freedom is obtained by Monte Carlo estimation with 300 simulations. The *SelectKBest* function in the python package *sklearn* is used for implementation.

Here in (5) the second equality holds because $E(\hat{y}_i)$ is a constant and could be pull out of the expectation, while $E(y_i - E(y_i))$ is 0. And in the approximation we generate $\mathbf{y}$ by the fixed $\mathbf{X}$ and fixed β for J times, compute the corresponding $\hat{\mathbf{y}}_j$ for each time, and $\bar{y}_i$ is the mean value taken with respect to j. In practical implementation, the order of division and summation could be swapped for convenience.

2.2 Numerical Simulation Examples

In this subsection, numerical simulations will be provided, as well as discussion.

2.2.1 Best Subset Selection

In 2, one could observe the non-monotonicity of the degree of freedom. The phenomenon could be studied by connecting best subset selection with so called "L-0 norm", which is not a real norm, and will introduce non-convexity into a optimization problem. Figure 6 in [2] demonstrates why non-convexity will lead to non-monotonicity of the degree of freedom, the key reason is that the projection of a point to a non-convex set could be unstable and not unique.

2.2.2 LASSO Regression

In [2], a specific example of LASSO is provided and it is reproduced in 3. The paper claim that it indicates non-monotonicity even in a convex constraint(L1 constraint). Although the explanation is insightful, the example itself could be flawed. The reversal of monotonicity happens at around 0.3, one could observe that at the corresponding region in the solution path, the coefficient of variable "1" is starting to get across 0, which is suspicious. Recall the closed form solution of LASSO

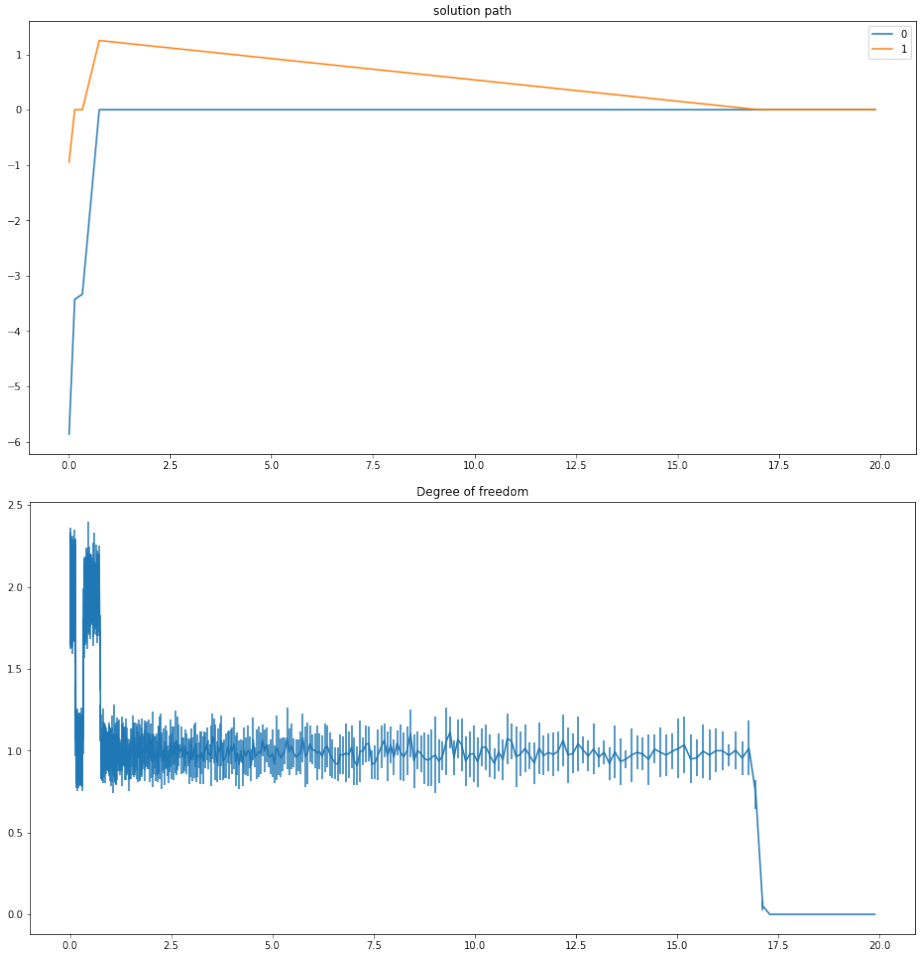

Figure 3: **Degree of Freedom and Solution Path of LASSO**: The data are generated samely as the setting in [2].

$$\mathbf{X} = \begin{bmatrix} 0 & 1 \\ 2 & -5 \end{bmatrix} \quad \beta = \begin{bmatrix} -6 \\ -1 \end{bmatrix} \quad \epsilon \sim N(0, 0.03^2)$$

The degree of freedom is obtained by Monte Carlo estimation with 1000 simulations divided into 10 batches to plot the error-bar, in order to make the plot tidy under large noise, as suggested in[2]. In the implementation, the python package *sklearn* is called to solve for LASSO.

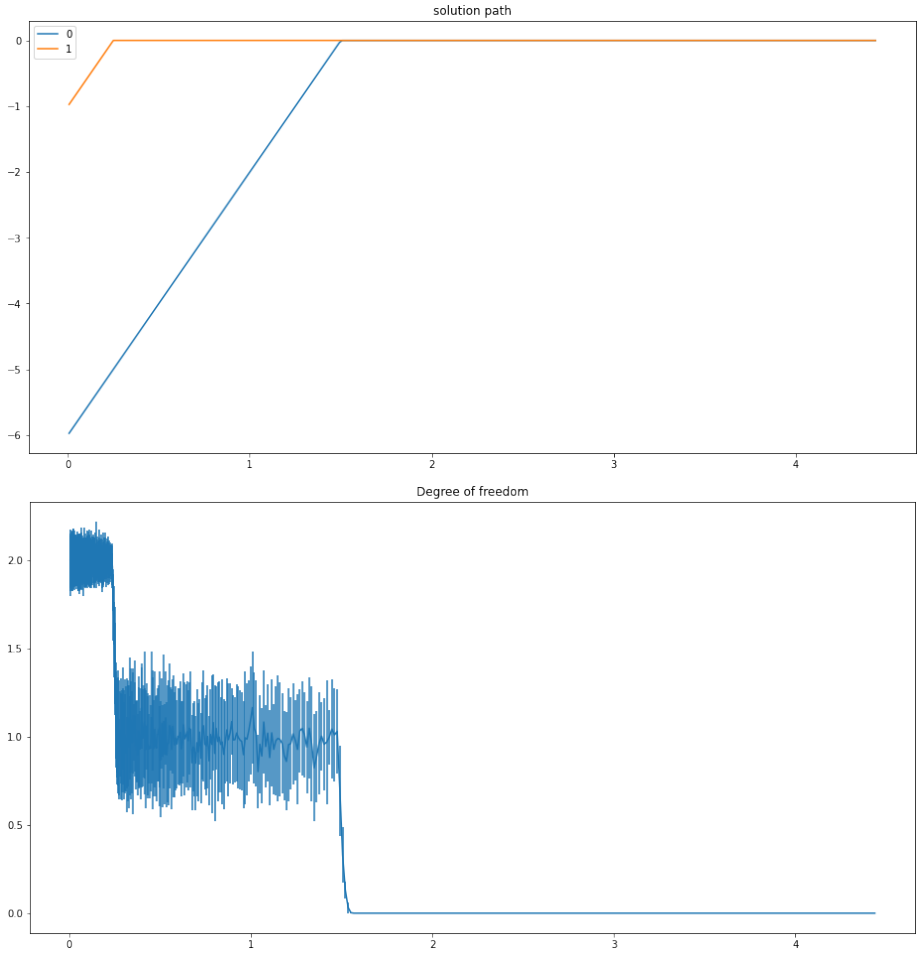

Figure 4: **Degree of Freedom and Solution Path of LASSO**: The data are generated samely as the setting in [2]. The degree of freedom is obtained by Monte Carlo estimation with 4000 simulations divided into 20 batches to plot the error-bar.

involving soft-shrinakge operator:

$$\beta_{LASSO} = \text{sgn}(\beta_{OLS}) \left(|\beta_{OLS}| - n_{sample}\lambda\right)^{+}$$
$$\text{or, } \beta_{i,LASSO} = \begin{cases} \beta_{i,OLS} + \lambda & \text{if } \beta_{i,OLS} \leq -\lambda \\ 0 & \text{if } -\lambda < \beta_{i,OLS} \leq \lambda \\ \beta_{i,OLS} - \lambda & \text{if } \beta_{i,OLS} > \lambda \end{cases} \tag{6}$$

The coefficient should not cross 0 in the solution path, according to the closed form solution.

With the implementation of the closed form solution, the result under same set-up is shown in 4. The reversal of monotonicity does not exist anymore. Hence it is likely that the unusual phenomenon of non-monocity in LASSO mentioned in [2] is caused by the approximation algorithm like gradient descent when solving LASSO. Nevertheless, the geometric explanation in [2] is still essential. Another graph 5 could be used to illustrate why even in a sequence of convex constraint set, the degree of freedom could still be non-monotonic.

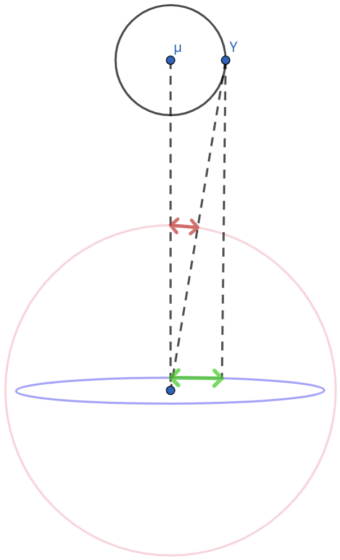

Figure 5: **Illustration of non-monotonicity in convex set sequence**: The black circle could be regarded as the distribution of $\mathbf{y}$, while the pink circle is a larger convex constraint set, containing a smaller convex set(the purple ellipsoid). The red arrow and the green arrow represent the covariance seperately, notice that in this case a smaller constraint set give a larger covariance.

2.2.3 Ridge Regression

The data are generated by a linear model with $n = 20, p = 5$. The result is shown in 6. The monotonicity of degree of freedom is quite clear, as mentioned in [2].

2.2.4 Simple Bivariate Model Revisit

The simple example provided in summary can be used to illustrate that the degree of freedom could be unbounded, both in simulation and in theory. The simulation result is shown in 7.

As for the theoretical part, the calculation in [2] skips several steps. Here a more detailed calculation is provided.

Set-up:

$$\begin{aligned} &\mathbf{y} = \mathbf{X}\beta + \epsilon \\ &\mathbf{X} = \mathbf{I}_2, \quad \beta = \begin{bmatrix} A \\ A \end{bmatrix}, \quad \epsilon \sim N(\mathbf{0}, \mathbf{I}_2) \end{aligned} \tag{7}$$

A best subset model of size 1 is applied to the 2-observation problem. Apply the second equation in (5):

$$\begin{aligned} DF(\beta, \mathbf{X}, FIT_1) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} E\left((y_i - E(y_i))\,(\hat{y}_i)\right) \\ &= \frac{1}{\sigma^2} \left[E\left((A + \epsilon_1 - A)\hat{y}_1\right) + E\left((A + \epsilon_2 - A)\hat{y}_2\right)\right] \\ &= \frac{1}{\sigma^2} \left[E\left(\epsilon_1 \hat{y}_1\right) + E\left(\epsilon_2 \hat{y}_2\right)\right] \end{aligned} \tag{8}$$

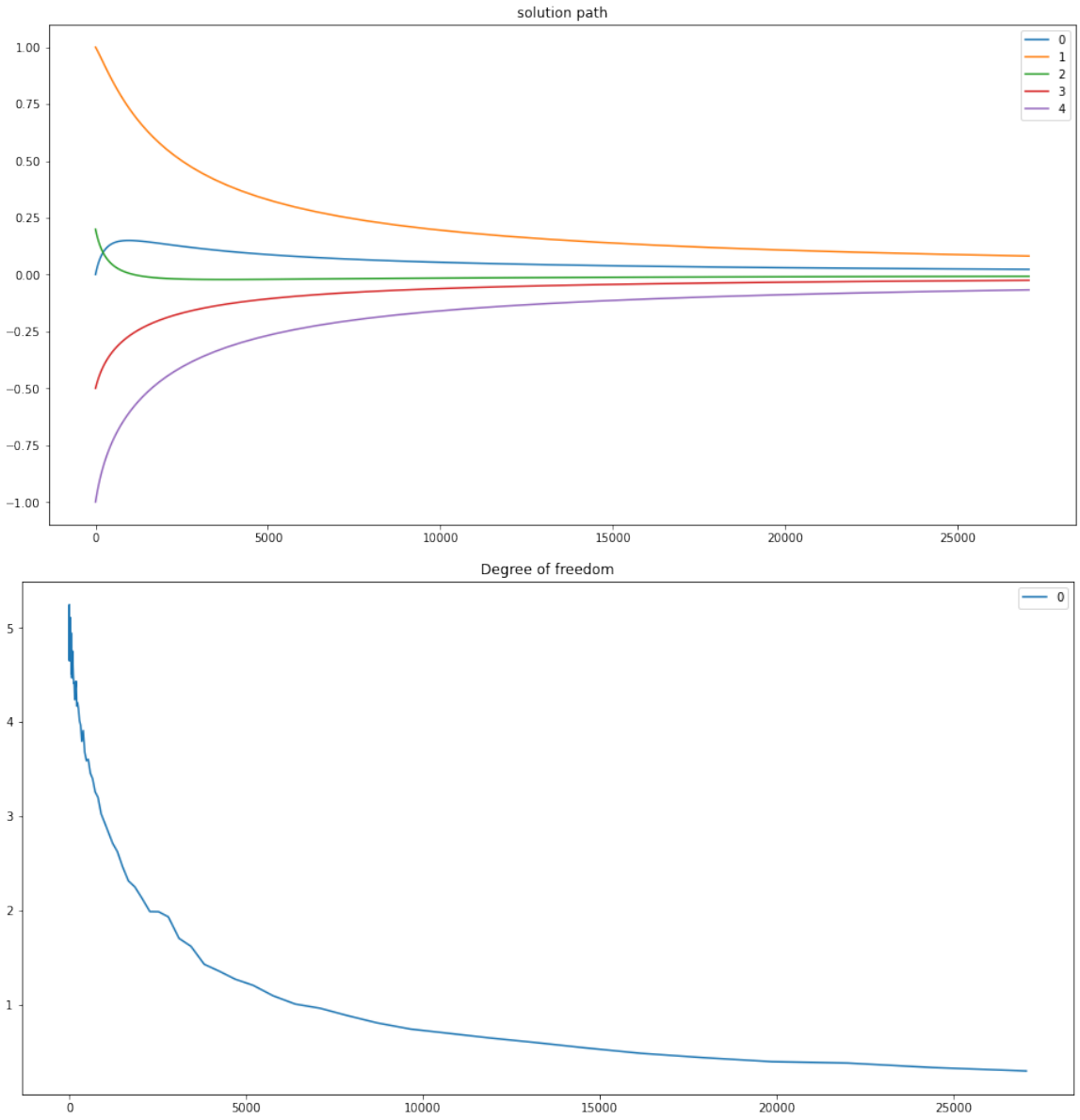

Figure 6: **Degree of Freedom and Solution Path of Ridge Regression.** The degree of freedom is obtained by Monte Carlo estimation with 500 simulations.

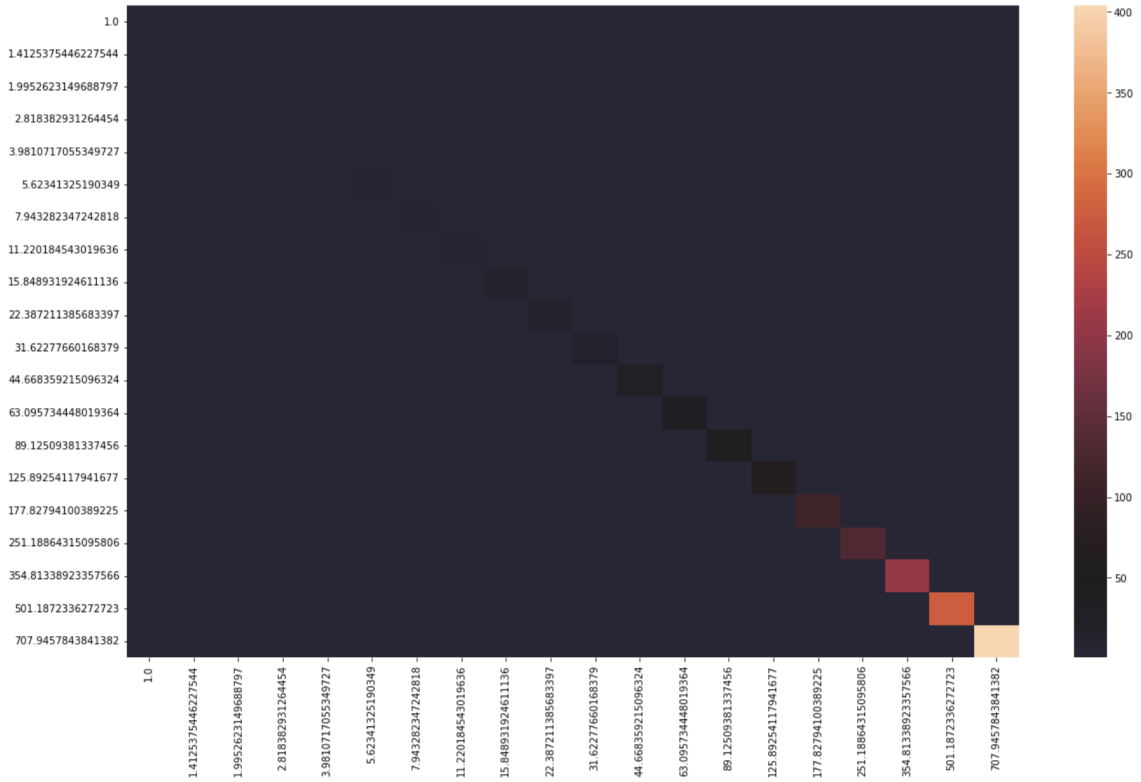

Figure 7: **Degree of Freedom for Best of Two Univariate Models**: The data are generated by a simple Gaussian bi-variate full model with $n = p = 2$. The design matrix X is identity. The axes represent the mean of y and the color in the heat map represents the degree of freedom achieved by a best subset univariate model. (Monte Carlo estimation with 300 simulations)

Notice that

$$\begin{aligned}\hat{y}_1 &= \mathbf{1}(\epsilon_1 > \epsilon_2) \cdot (A + \epsilon_1) + \mathbf{1}(\epsilon_1 < \epsilon_2) \cdot 0 = \mathbf{1}(\epsilon_1 > \epsilon_2) \cdot (A + \epsilon_1) \\ \hat{y}_2 &= \mathbf{1}(\epsilon_2 > \epsilon_1) \cdot (A + \epsilon_2)\end{aligned} \tag{9}$$

Hence

$$\begin{aligned}DF(\beta, \mathbf{X}, FIT_1) &= \frac{1}{\sigma^2}\left[E\left(\epsilon_1 \mathbf{1}(\epsilon_1 > \epsilon_2) \cdot (A + \epsilon_1)\right) + E\left(\epsilon_2 \mathbf{1}(\epsilon_2 > \epsilon_1) \cdot (A + \epsilon_2)\right)\right] \\ &= \frac{1}{\sigma^2}\left[E(\max(\epsilon_1, \epsilon_2))A + E(\max(\epsilon_1, \epsilon_2)^2)\right]\end{aligned} \tag{10}$$

In the simulation, $A \approx 707.9458$, plug in $\sigma = 1$ and $E(\max(\epsilon_1, \epsilon_2)) \approx 0.5642$, the result is consistent with $DF \approx 400$ in the simulation.

3 Conclusion

To explore more about the concept of degree of freedom, this report reproduce the numerical simulations in [2], which reveal a few irregularity of degree of freedom. In the simulations, the degree of freedom could be non-monotone or unbounded. Other simulations like the Ridge Regression and LASSO with closed form solution implementation are also provided.
Moreover, [2] introduced a geometric view as theoretical explanation so that the concept of regularization is associated with constrained feasible sets. Monotonicity could be defined by inclusion relationship of sets in such view. Further more, the irregularity of sets(for example, non-convexity) could partially account for the irregularity of degree of freedom. A graph(5) is shown for illustration.
In conclusion, though the definition of degree of freedom indeed captures the correlation between in-sample error and out-of-sample error, one should still beware of other intuitions of DF that generalized from the standard linear regression.

References

[1] Efron, Bradley. "How biased is the apparent error rate of a prediction rule?." Journal of the American statistical Association 81.394 (1986): 461-470.

[2] Janson, Lucas, William Fithian, and Trevor J. Hastie. "Effective degrees of freedom: a flawed metaphor." Biometrika 102.2 (2015): 479-485.

[3] Mallows, C. L. "Some Comments on CP." Technometrics, vol. 15, no. 4, 1973, pp. 661–75. JSTOR, https://doi.org/10.2307/1267380. Accessed 14 Dec. 2022.