Empirics: Virtue of Complexity

Kanji

Last update: 2024-07-11

Contents

This document describes the results of the empirical experiment on the virtue of complexity using random features, as conducted in Kelly et al. (2024). As substantiated in Section 5 of Kelly et al. (2024), the Sharpe Ratio is expected to exhibit a double-ascent under some assumptions. The performance of the corresponding kernel in the limit (Neural Network Gaussian Process) is also reported. This is computationally feasible because the number of raw features is only fifteen.

The hyperparameters are listed in the following table:

Parameters	Values
γ of RFF	2
Number of Simulation	500
Maximum Number of Observable Features	600
Window Size	12
Ridge Regularization	1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3

Data is retrieved from the Replication Code of Kelly et al. (2024). All variables are scaled in the same manner with them.

In Section 1, the Random Fourier Feature is examined. In addition to the Normal weights, other distributions (Laplacian and Caucy) for weights are examined.

In Section 2, alternative activation functions (ReLU, Error Function) are examined.

Variations in the Random Weights

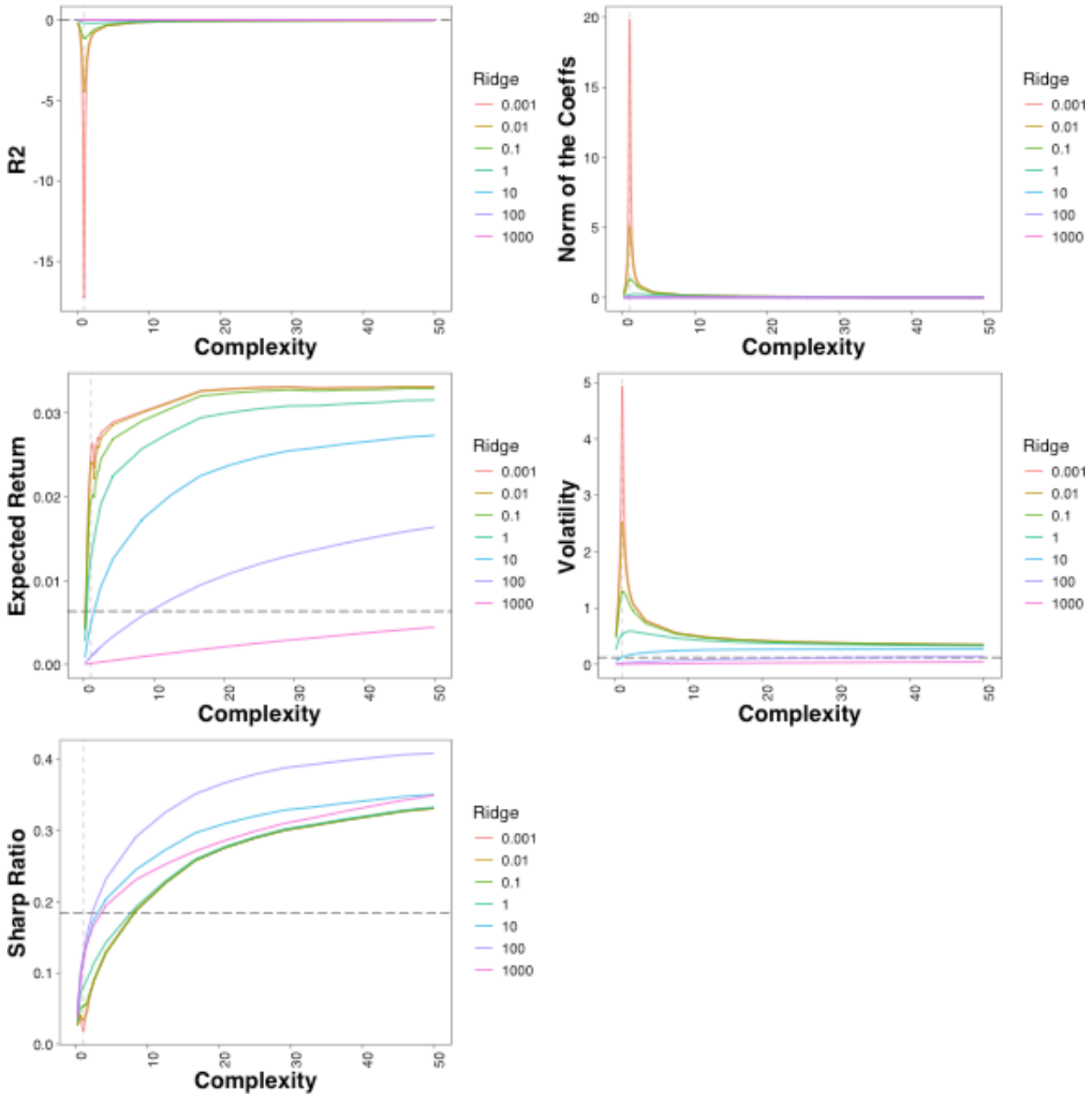
Generate the random fourier feature with random weights of Gaussian, Laplacian and Caucy following Rahimi and Recht (2007). The grey dotted line represents $c = 1$. The black dash line represents the performance of the kernel regression with corresponding NNGP.

As can be seen, the increasing pattern in the out-of-sample market-timing Sharpe ratio is observed in Laplacian and Caucy as well as in Gaussian. This is consistent with the misspecified model in Kelly et al. (2024). The volatility and R^2 converge to those of the kernel regression as expected. As for the mean

and Sharp ratio, we observe deviations between the Neural Network with wide width and the corresponding kernel. The potential reason for this is the estimation error or insufficient number of samples to evaluate the statistics.

Gaussian

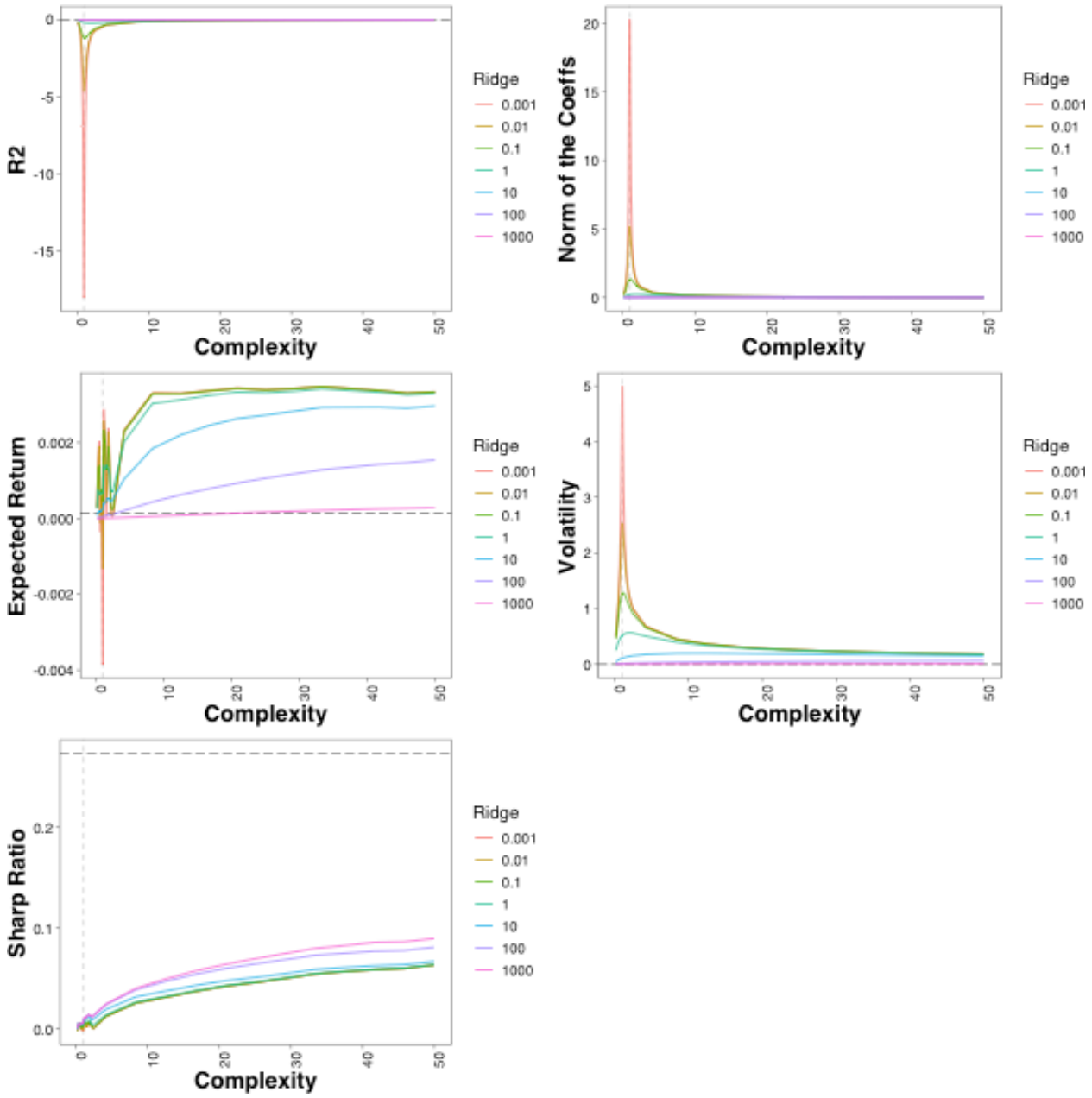
```
g <- plot_stats(df_stats, df_kernel_stats)
```




```
#
```

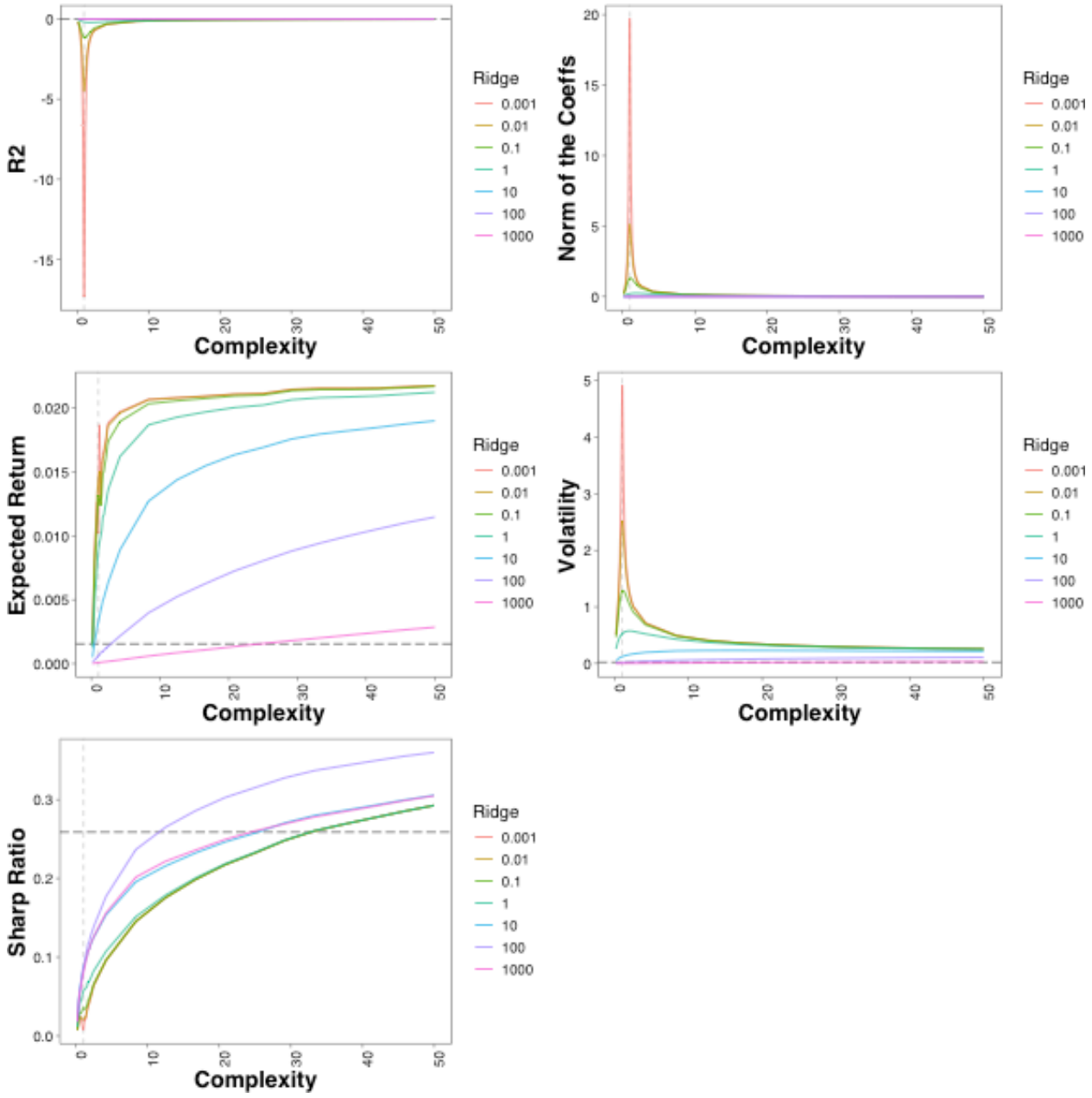
Laplacian

```
g <- plot_stats(df_stats, df_kernel_stats)
```



Caucy

```
g <- plot_stats(df_stats, df_kernel_stats)
```



Variations in the Activation Function

Alternative activation functions (ReLU and the Error Function) are examined. As the NNGP theory substantiated, the Neural Networl converges to the kernel in the limit of infinite width in some cases.

In ReLU case, the corresponding NNGP is

$$K(x_1, x_2) = \frac{1}{\pi}\left(x_1^\top x_2\left(\pi - \arccos\left(\frac{x_1^\top x_2}{\|x_1\|_2\|x_2\|_2}\right)\right) + \sqrt{\|x_1\|_2^2\|x_2\|_2^2 - (x_1^\top x_2)^2}\right) \tag{1}$$

In erf case, the corresponding NNGP is reported in Williams 1996 as

$$K(x_1, x_2) = \frac{2}{\pi} \arcsin\left(\frac{2x_1^\top x_2}{\sqrt{1 + 2x_1^\top x_1} \cdot \sqrt{1 + 2x_2^\top x_2}} \right) \tag{2}$$

The grey dotted line represents $c = 1$. The black dash line represents the performance of the kernel regression with corresponding NNGP. To improve the stability, I increase the number of simulation to 1000. To omit the extreme values, the mean across samples is computed after truncating at the upper and lower 99th percentiles.

In ReLU case, R^2 monotonically decreases along with the model complexity. The norm of the coefficient converges to a value close to zero because the ridge regression achieves the minimum l_2 norm. The expected mean is negative when c is greater than 1 in the ridgeless case. Thus, in ReLU case, the variance is greater than the bias. Moreover, the high ridge regularization achieves positive mean and Sharp ratio. This further supports that the variance reduction is important even it damages the bias.

We do not observe the double-ascent property in the ReLU case. Indeed, the Sharp ratio decreases as the model becomes complex. Moreover, the expected mean decreases when c is greater than 1 in the ridgeless case. This looks similar to the theoretical pattern in the correctly specified case of Kelly et al. (2024). Thus, I derive this to three potential violations of assumptions.

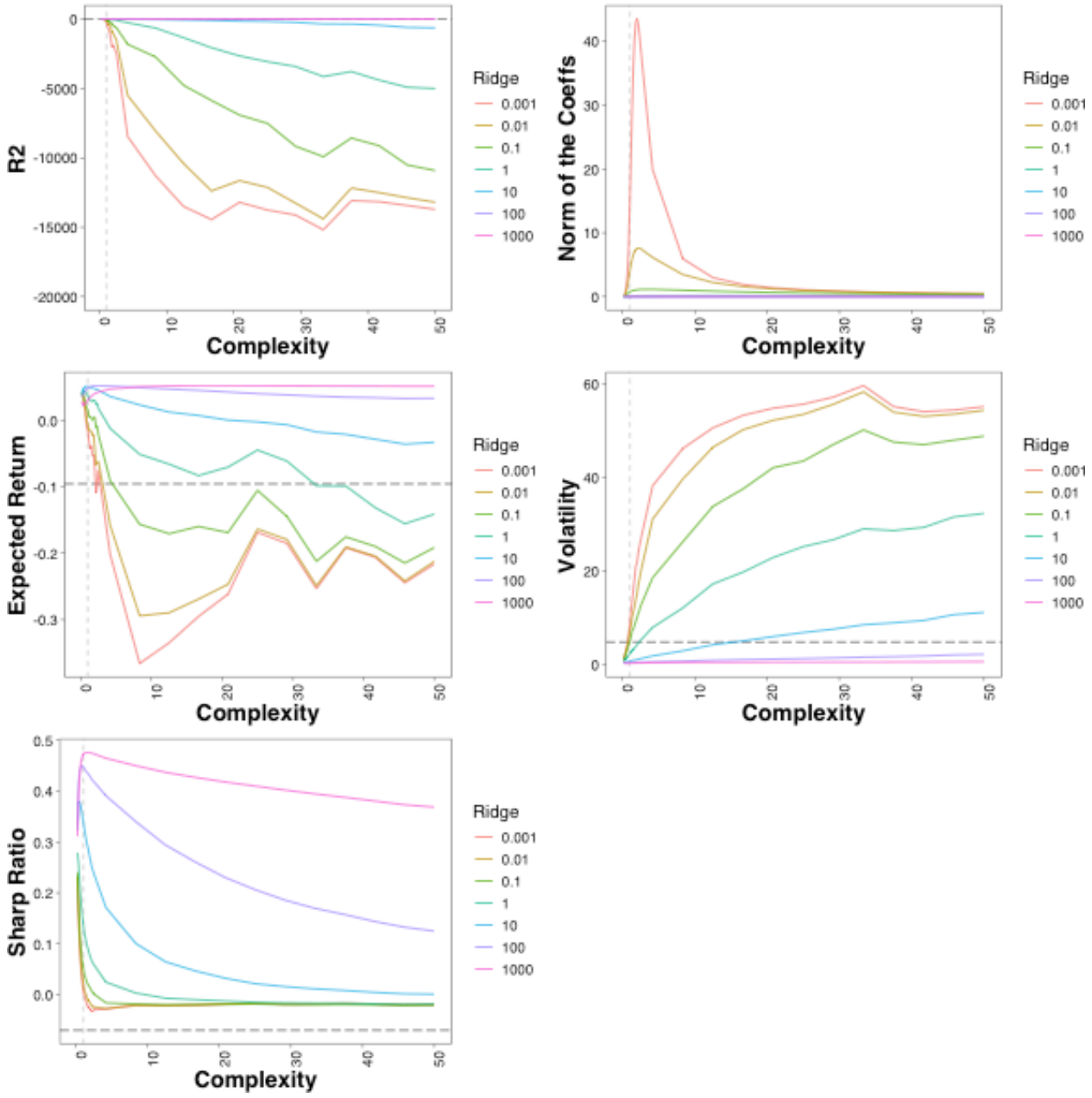
- The random features are not part of the true Data Generating Process (DGP). The Gaussian, Laplacian and Caucy kernel may be closer to the DGP than NNGP of ReLU and erf.
- The map from the independent random vectors to the return is nonlinear. This case is discussed in Hastie et al., (2022).
- The true DGP has only finite number of factors, which results in the specified case.

The Sharp ratio achieved by the kernel in the ReLU case is negative. This supports the first and second hypotheses that the true DGP is irrelevant of the kernel corresponding to ReLU.

In the erf case, we observe that the high complexity and regularization imporve the R^2 and Sharp ratio. This result is consistent with the misspecified model in Kelly et al. (2024).

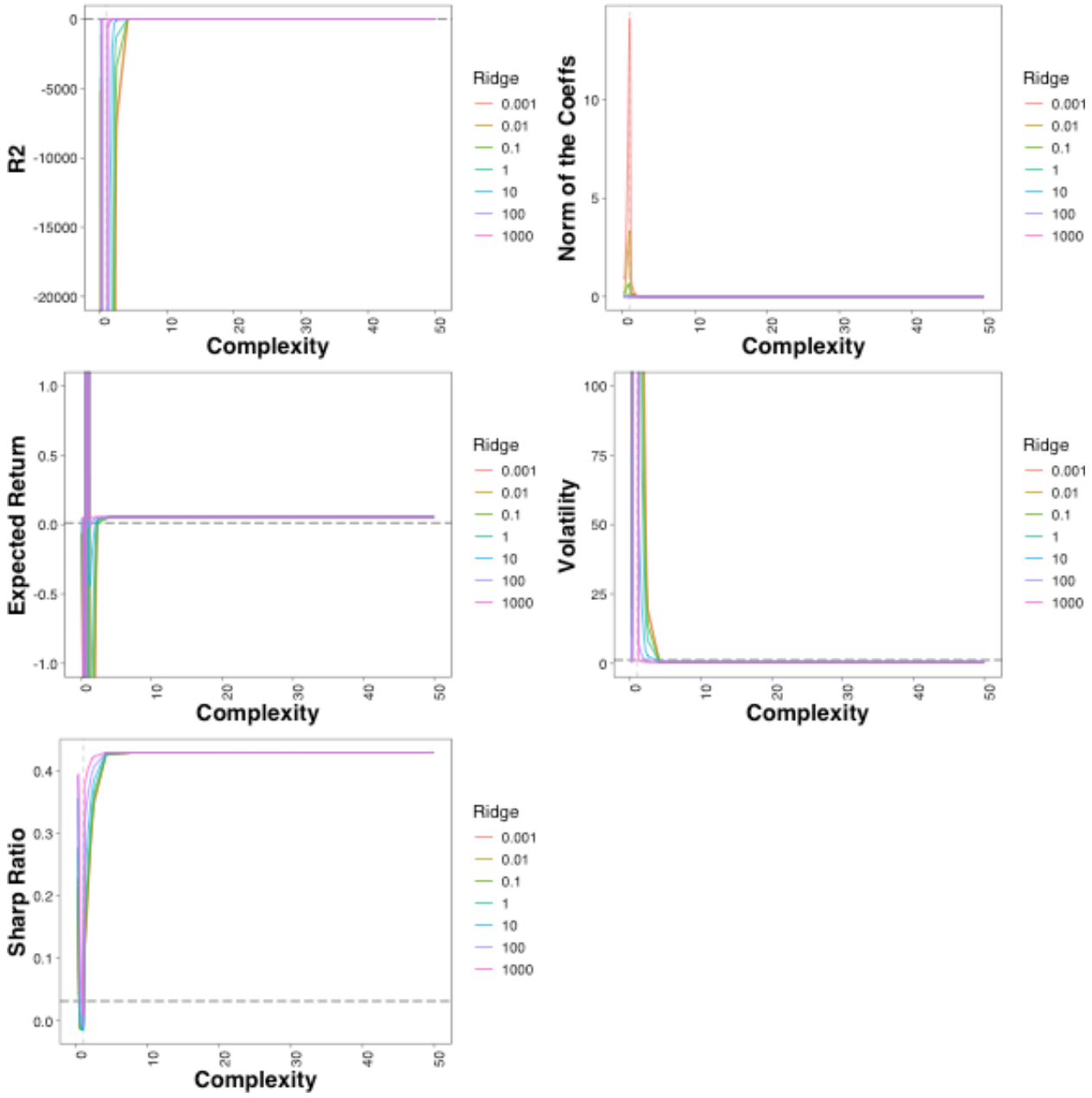
ReLU acitvation

```
g <- plot_stats(df_stats, df_kernel_stats)
```



Error Function acitvation

```
g <- plot_stats(df_stats, df_kernel_stats)
```


R2
0
-5000
-10000
-15000
-20000
0
10
20
30
40
50
Complexity
Ridge
0.001
0.01
0.1
1
10
100
1000
Norm of the Coeffs
10
5
0
Complexity
Ridge
0.001
0.01
0.1
1
10
100
1000
Expected Return
1.0
0.5
0.0
-0.5
-1.0
Complexity
Ridge
0.001
0.01
0.1
1
10
100
1000
Volatility
100
75
50
25
0
Complexity
Ridge
0.001
0.01
0.1
1
10
100
1000
Sharp Ratio
0.4
0.3
0.2
0.1
0.0
Complexity
Ridge
0.001
0.01
0.1
1
10
100
1000