Bagging Linear Predictions(Problem 2)

Mohamad Lakkis

May, 2024

This PDF is to explain some of the coding part in the *Bagging Linear Predictions (Problem 2)*

General Notes

- In python there is two codes, they are both the same except for the last part, one is actually done via parallel processing and the other is done via a for loop. (This is to make the running process faster)

Proving that $Var(f(X)) = \beta^T \Sigma \beta$

This is very trivial to prove, since we know that $X \sim N_p(0, \Sigma)$, and we have that $f(X) = X\beta$ (Of course the variance is with respect to X!)

$$Var(f(X)) = Var(X\beta) = \beta^T Cov(X)\beta = \beta^T \Sigma \beta$$

Analysis

General Analysis

- We have two figures, the first is when we have used 5 features in the actual model, (i.e. $f(X)$ contains 5 features of X). The second figure is when we have used 25 features in the actual model.
- Firstly, we can see that in all of them we have approximately the same trend, the bagging method is better than the regular until a certain threshold(should be close to p), then the regular method is better.
- Regular FWS performs better after a certain threshold because it explicitly avoids selecting redundant features (since it is a greedy method, select the best choice with what it has), while bagging FWS may start following the noise and the correlation between data when the number of features selected increases. The question remains why this threshold is approximately the same(between 22 and 25 number of features) for all the plots, which I didn't know the answer to.

- So what we can say is that generally, bagging works well when fewer features are required, by as we know reducing prediction variance. But, as number of features increases, and particularyly after a certain threshold(22-25 in this case) bagging FWS will start following irrelavant correlation leading to a higher variance, while regular FWS remains more stricted on how it is selecting these features.

- The main purpose of bagging as we know, is to reduce variance by combining models(which we can see does a very good job with low number of features), which after some point will lead to huge errors in the prediction, and after this point the Regular FWS will have a lower variance than the bagged one since it is accumaling more and more features leading to a less window for variance.

Comparison row wise

- For the first row (where we are using in both cases 5 features and $\rho = 0.3$, but differing with the SNR value), we can see that both graphs have the same shape, and both of the curves reach the minimum at the same point, which is around 5 (which is logical since we have used in the true f 5 features). But we can see that bagging performs better, and that is because regular FWS have much variance, where bagging tries to lower this variance, which it was successful in doing. One thing to notice that yes both graphs have the same shape, but the one on the right have better errors, and that is logical since we have a better SNR value, which means that the noise is less, which makes the prediction more accurate.

- But now if we look at the second row, where the same environment is as one, but $\rho = 0.8$, there is no minimum point, meaning that since the data is highly correlated, both methods will start following the correlation, but the bagging method will be slightly better, since it tries to reduce the variance. (of course we are talking until the threshold that we already talked about). And as before even though the shape is the same, the one on the right is better, since the noise is less (i.e. SNR higher).

- Same Analysis applies for the last two rows

Comparison column wise

- What we mean here is we will compare for each figure the two columns. So we want to see how ρ affect it.

- We can easily see that as expected with a higher ρ, a higher correlation between the data which will lead to a higher error, and that is what we can see in the figures. But notice that also here the effect of the number of features used(in the true f) plays a role, we can see that since we have increased the number of features used to 25, and increased the correlation between them, the model will a have a very hard time in selecting the right features, and that is why we can see that the error is increasing as we increase the number of features used. And we can see the values of the errors are relatively to the first figure huge.

- Note that in set-up 6, the same shape is applied but it is because at a high number of features, we can see how the bagged model has high variance which can lead to this huge jump in the error. so as we increase N how many times we are repeating the experiment we can expect(which I tried) that this form will be in most of the graphs, especially in the graphs where we have high correlation between the features. (because of the high variance of the bagging when number of features increases close to p)

Conclusion

- We can clearly that the correlation between the features is the one that affect our accuracy the most, if we have high correlation between the data we will get huge errors.

- We can see that the bagging method is better when we are using a smaller number of features, since it reduces the variance (which will be very high in the case of regular FWS), but of course as we get closer to p, the FWS will have lower variance where as the bagging method will start having a higher variance compared to the regular FWS.

Plots and Graphs

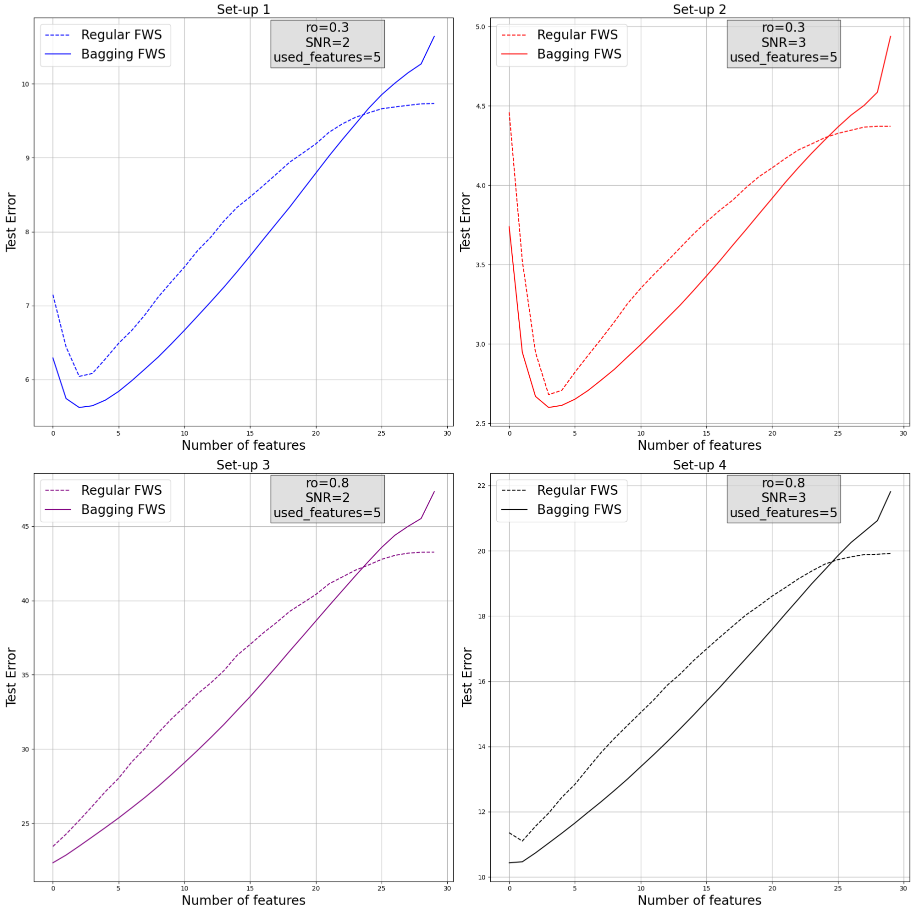

Figure 1: Number of features used = 5

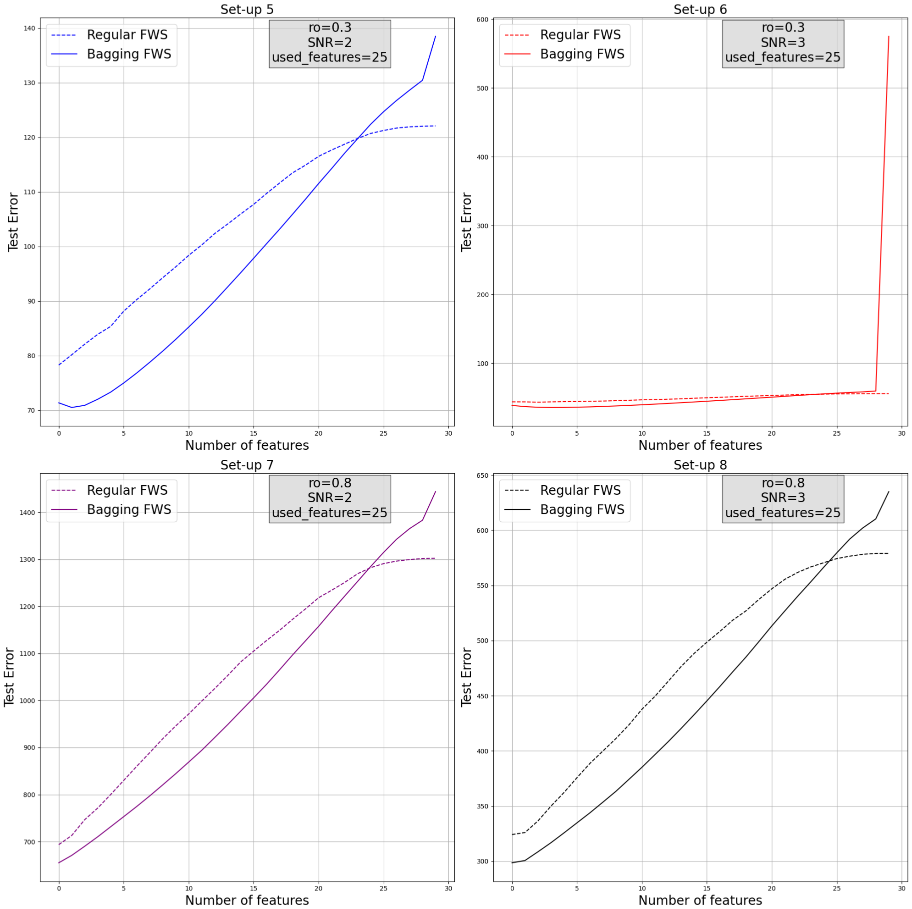

Figure 2: Number of features used = 25