Numerai
Competition Analysis
and Improvement Proposals

Anonymous Contributors

August 2017

1 Introduction

This document summarizes our main findings after a preliminary analysis of the originality and concordance scores, as well as some other aspects of the competition. As a result, we propose some simple ways of addressing the main problems that have been detected, and suggest further improvements related to the data set and the stacking competition.

2 On the Concordance

One of the characteristics required for a submission to be eligible for winning the classical and stacking tournaments consists on the concordance among the predictions on the Validation, Test, and Live sets. The concordance refers to the statistical distribution of the predictions, which should be similar in the three referred sets. The actual algorithm (as of 7/7/17) is described below:

Algorithm 1 Pseudocode for the Concordance Algorithm

```
All the data (Training and Tournament) is divided into $K = 5$ clusters by means of the K-Means algorithm.
for Each Cluster do
    Compute the maximum of the Two-Sample Kolmogorov-Smirnov score (deviation from the hypothesis of equal distributions) for the pairs Validation-Test, Validation-Live, and Test-Live.
end for
Obtain the final Concordance Score as the average of the $K = 5$ available measures.
Declare concordance if the final score is lower than a given threshold (0.12).
```

As can be seen, the main ingredients of the current Concordance Algorithm consist in the use of the Kolmogorov-Smirnov (KS) test, and a preprocessing stage based on the K-Means clustering algorithm.

2.1 Main Problems with the Concordance Test

Although reasonable, the current test for concordance exhibits several drawbacks that should be addressed in the near future. In the following, we summarize what we have identified as the main problems.

The Clustering Stage

First of all, the clustering stage based on K-Means seems completely unnecessary. We suppose that this stage was introduced at some point in order to correct some kind of problem. However, one should be aware that the accuracy of the KS test, and therefore also of the final Concordance Score, is limited by the sample size in the live set (around 1500 samples), and even assuming equal-size clusters (which is not the case in our experiments, see Figure 1), the effective sample size reduces to 300, which could be considered as too low to get accurate results.

Additionally, the selection of the K-Means algorithm is implicitly introducing a preference for the Euclidean distance among the different data points, and it could somehow interfere with some prediction algorithms. We ignore whether such choice has anything to do with the structure in the data, but if it has, this could be considered as a kind of (not too dangerous) data leakage. On the other hand, if the choice has nothing to do with the data structure, and the only purpose consists in splitting the tournament data, maybe a random partition of the tournament points would be preferable.

Finally, we must point out that the non-stationarity of the data should be taken into account in the clustering process. In other words, the clustering stage is grouping points exclusively in base to their Euclidean distances, whereas a quick analysis of the data shows a clear trend on the per-epoch means and standard deviations of several features, as is shown if Figure 2.

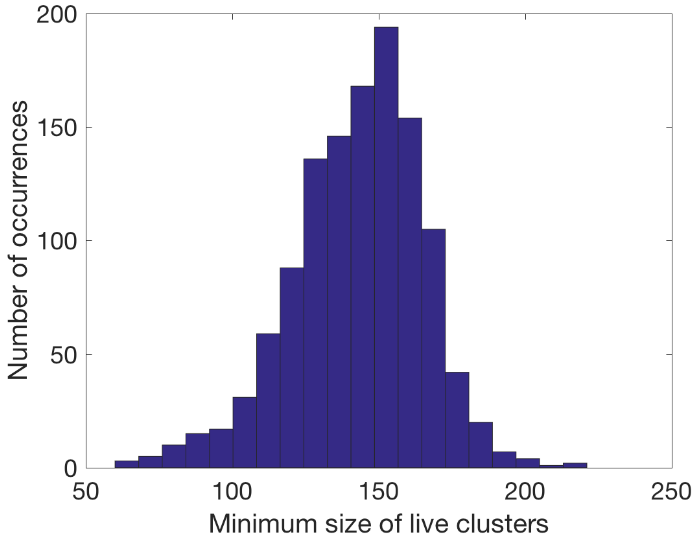

Figure 1: The live set size in tournament 65 contains 1259 samples, and by doing 5 clusters it could be expected to have 250 nominal samples per cluster. We have performed 500 simulations of the clustering and recorded the size of the smallest one of the clusters of the live set. In this figure, it can be seen the variability of the minimum cluster size and the high probability of obtaining clusters smaller than 100 samples.

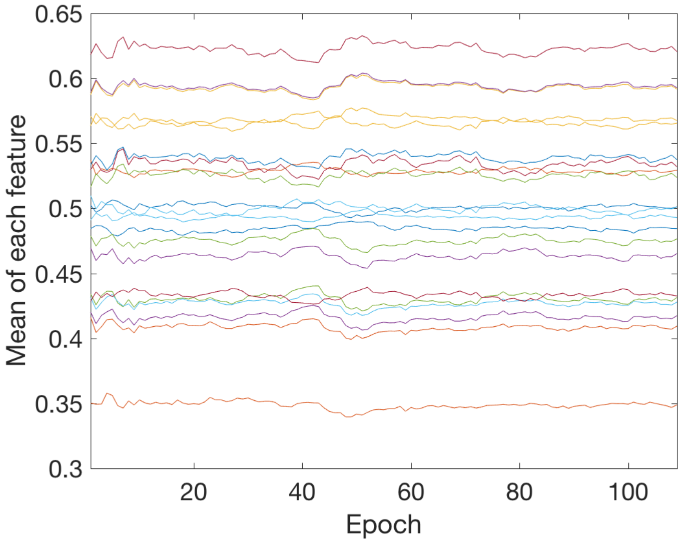

Figure 2: This figure illustrates the non-stationarity of the data, which should be taken into account in the clustering process. A change in the mean of the features can be observed around epoch 45.

Combination of Two-Sample KS Tests

From our point of view, the use of the Two-Sample KS test is suboptimal. Firstly, although previously proposed in the literature [1, 2], the use of the maximum of the three KS scores is completely heuristic. More importantly, this “fusion rule” based on the maximum of the three scores, does not take into account the very different sample sizes of the Validation, Test, and Live sets, and it is therefore dominated by the small sample size of the Live Set.

Additionally, the threshold selection is a far from trivial problem because the false alarm (declaring lack of concordance when the predictions come from the same model) probability does not only depend on the threshold, but also on the underlying distributions of the predictions. This is illustrated in Figure 3, which clearly shows that the same threshold results into different false alarm probabilities for different distributions of the predictions.

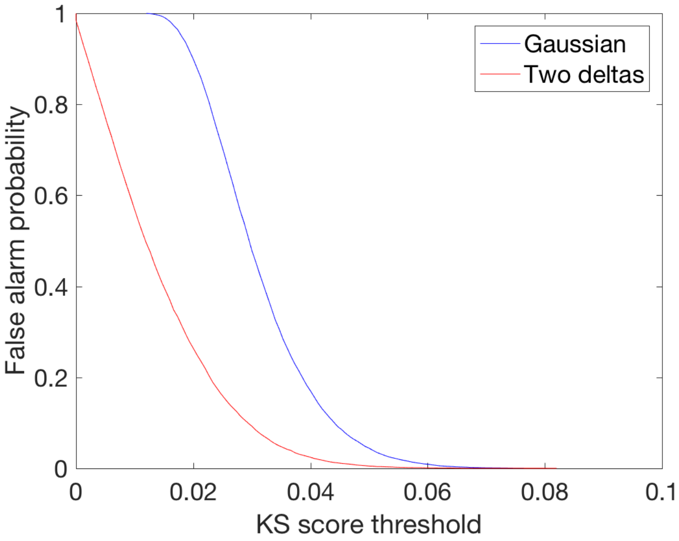

Figure 3: The threshold selection of the KS score affects the probability of false alarm, and as this figure illustrates, it also depends on the probability distribution. As an example, two distributions has been simulated: a Gaussian of mean 0.5 and standard deviation 0.1, and one consisting of two Dirac deltas at 0.4 and 0.6. As can be seen, the false alarm probability curves differ significatively.

Test Set Labels

Finally, we are sure that the Numerai team can clearly detect (many?) submissions which, even passing the concordance test, are clearly non concordant. The use of the labels in the test set will clearly identify submissions with very different logloss in the Validation and Test sets. Of course, a naive use of the test labels to compute the concordance score would easily result into a leakage of information, but this is a possibility that should not be discarded.

2.2 Proposed Solutions to the Concordance Issues

Based on the previous analysis, we have identified the following corrective actions in order to improve the performance of the Concordance Test:

1. **Avoid the Clustering Stage**: We propose to reconsider the Concordance Test from the scratch, removing the clustering stage. If needed, the clustering or other alternative approach could be introduced at a later stage. However, we strongly believe that any problem addressed by means of the clustering approach should be revisited with a more theoretically solid point of view.

2. **Design of Alternative K-Sample Tests**: The drawbacks of the KS Test could be avoided by revisiting the testing problem from first principles. This includes the design of K-sample tests for populations with different sample sizes, as well as the use of practical methods for the selection of the threshold, focusing on techniques invariant to the underlying null distribution.

3. **Reconsider the use of the Targets in the Test Set**: The comparison of the logloss in validation and test sets would be a definitive measure of concordance. However, this would incur into a data leak that could be exploited to improve the predictions of the test set. As a preliminary idea for avoiding this information leakage, one could evaluate the logloss in the test set by using a random (for each submission) subset of the test set data.

3 On the Originality

In this section we examine the procedure for checking the originality of one submission. At the current time (7/7/17), the originality analysis is as follows:

Algorithm 2 Pseudocode for the Originality Algorithm

```
Initialize the submission as original
Compute the correlation coefficient and the Two-Sample KS score between the submission and
  the most recent submissions of the remaining competitors.
if The largest correlation coefficient is lower than a threshold (0.95) then
    The submission is not original.
end if
for Different K values do
    if The K-th largest KS score is lower than a threshold $\mu_K$ then
        The submission is not original.
    end if
end for
```

The originality check is therefore based on the correlation coefficient computed from the tournament data predictions, and (surprisingly) from the KS scores. In its current (7/7/17) implementation, the algorithm uses the values $K = 1$ and $\mu_K = 0.03$.

3.1 Main Problems with the Originality Test

Let us point out here what we think are the main problems of the current Originality Test.

The Use of the KS Test

The Originality Test should be... a originality test, not a distribution test. The competitors are already investing a lot of time trying to get original predictions. It seems excessive to also ask them to provide original distributions. In other words, it seems that with the current implementation, two submissions independently generated at random from the same distribution would be qualified as non original, even when they are completely independent.

Obviously, this problem translates into the fact that the time at which the submissions are made plays a determinant role in the result of the competition, providing a clear advantage to those competitors making their submissions in the first hours, even when these users are not even looking at the data. An analogous interpretation would be that once a competitor has a submission declared original, he/she has become the ”owner” of a statistical distribution, which seems really unfair.

Finally, let us point out that, although it makes sense for the analysis of concordance, the use of the KS test is not recommended for checking originality because, apart from the previous reasons, a submission can be easily modified in order to get a significant boost on the KS score. For instance, a simple quantization (with randomly generated levels) of the predictions, can result into a much greater KS score.

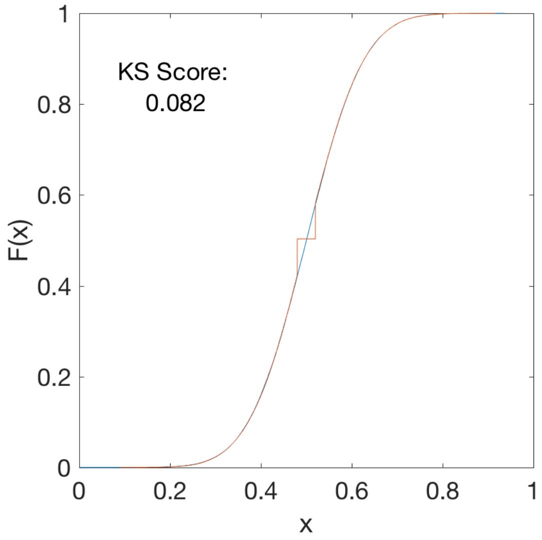

Figure 4: When used for checking originality, the KS test can be easily cheated. In this simulation, we have obtained two realizations of a Gaussian distribution of mean 0.5 and variance 0.1. In one of the realizations, we have truncated to 0.48 all the values between 0.48 and 0.5, and change to 0.52 all the values between 0.5 and 0.52. With this simple trick, the KS Score between the original and tweaked submissions is 0.082, which allows to pass the originality check.

Testing for Independence

The current implementation of the Originality Test looks for linear correlations. However, we think that a more general independence test would result into more reliable results. Additionally, we must point out that with the current implementation, and assuming that the distribution of the predictions is symmetric around 0.5, two identical submissions could be made to appear as original by just replacing some of the scores (s) in the Test Set of one of the submissions, by their complementary $1 - s$.

Other Minor Issues

For completeness, let us point out an additional couple of issues:

- In the comments of the provided code, it seems that the current submission is going to be compared with the previous one and, under some situations, the originality score will be propagated. Although this is not finally included in the code, we would like to point out that this kind of idea should be avoided, especially if, as we suppose, the originality code is going to be made available to all the competitors.
- As many users have requested, it would be nice to have some indicator in the web page in order to know when the originality has been calculated. In any case, the long delays in the computation of the originality do not seem to be justified by the complexity of the code.

3.2 Proposed Solutions to the Originality Issues

Based on the previous analysis, we suggest the following changes in the Originality Test:

1. **Avoid the KS Tests**: As previously pointed out, the distribution of the predictions has (almost) nothing to do with their originality. Moreover, the KS test can be easily “cheated” as previously described. Therefore, we strongly recommend to avoid these tests.

2. **Independence Tests**: The current test for correlation should be replaced by a more general and robust test for independence. Moreover, we suggest to calculate (possibly different) independence scores in the Validation, Test and Live sets. In this manner, we could avoid ad-hoc modifications of the Test Set predictions in order to achieve originality.

4 On the Data Set and the Stacking Tournament

Let us conclude this document by pointing out three more general issues related to the Numerai competition:

The first one relates to the Data Set, and the apparently random results pointed out by some competitors in the Slack channel. Although we believe in the predictive power of the competitors submissions, we agree that the small size of the Live Set makes really difficult to provide a set of predictions consistently beating the best of, lets say 200 random submissions. The obvious solution to this problem would consist in a larger Live Set, but we guess this is not a feasible solution due to several reasons. However, we have been wondering whether some kind of data augmentation for the Live Set would be possible. That is, we suggest to generate additional live data, maybe from simple combinations of the rows, encrypt them with the same keys/algorithms, and use this augmented Live Set to obtain more accurate estimates of the competitor’s models.

The second issue consists in the absence of Era information in the Tournament Data. Our data analysis suggest that the Live Set corresponds to one Era, and we conjecture that the Test Set consists of several Eras. In any case, the Era information in the Training Set seems to be really helpful in order to make predictions, and we can not think in any disadvantage for Numerai if the Era information of the Tournament Data is released.

The third issue relates to the Stacking Tournament. Apart form a careful analysis of the relationship among the evolution of the NMR price, its volatility, and the stacking behaviour of the competitors, we have been wondering whether it makes much sense to use different evaluation metrics for the Classical and Stacking competitions. From a competitors point of view, the design of a model for the Classical or Stacking competition would be significantly different. An alternative to the current evaluation of the Stacking competition, making it comparable to the Classical one, would consist in establishing a fix “confidence” value for all the users, and allowing them to select the performance threshold (currently log(2)) to be achieved. As an example, and with a constant confidence factor of 0.1, a competitor could decide to bet 10 NMR with an objective of 0.691. Thus, if his/her bet is selected (starting by those with the lowest objective) he/she will win 100 USD if the logloss in the Live Set is lower than 0.691, and the stacked NMR will be destroyed otherwise.

5 Conclusions

This first report has pointed out several drawbacks of the current Concordance and Originality Tests and has proposed several potential solutions to the main problems. We hope that some of these ideas will be helpful to improve the Numerai Tournaments.

Bibliography

[1] Herbert T. David. A Three-Sample Kolmogorov-Smirnov Test. *Ann. Math. Statist.*, 29(3):842–851, 09 1958.

[2] J. Kiefer. K-Sample Analogues of the Kolmogorov-Smirnov and Cramer-V. Mises Tests. *Ann. Math. Statist.*, 30(2):420–447, 06 1959.