STA 325 - Writeup

Ethan Shen, Austin Jia, Malavi Ravindran, Steven Herrera

11/1/2019

Introduction

Fluid motion is an incredibly complex topic which has significant applications in fields such as engineering, astrophysics, and climatology. Of particular importance is the concept of turbulence which, though easily observable, has been termed the "last great unsolved problem in classical physics." Metrics such as the Reynolds number (`Re`), which measures the intensity of a turbulent flow, Froud number (`Fr`), which quantifies gravitational acceleration, and Stokes number (`St`), which describes the size density of particles, can be used to bolster understanding of the particles in turbulence.

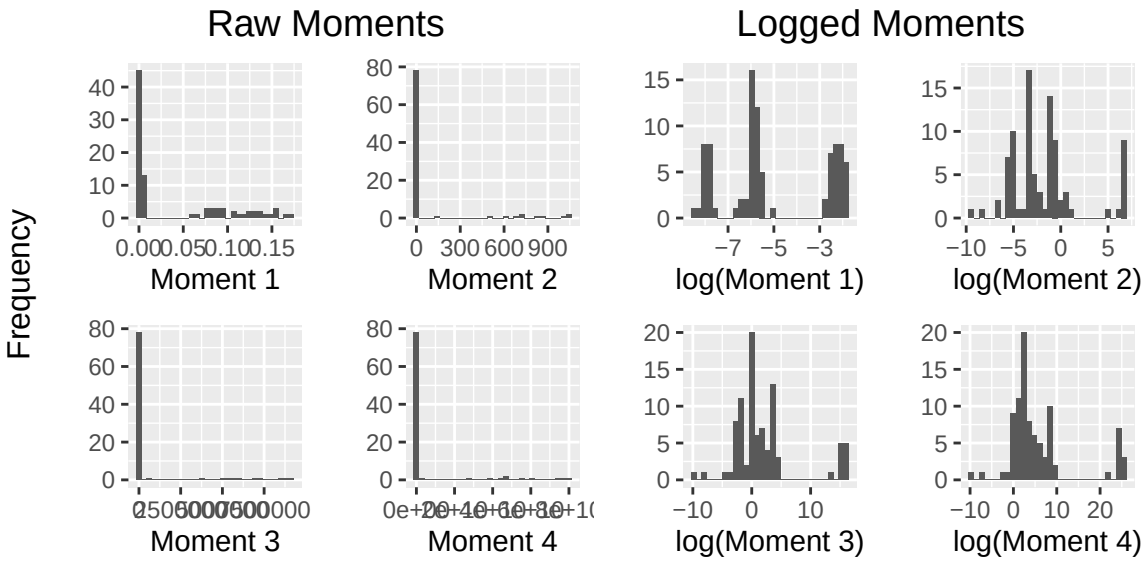
Our goal throughout this data expedition was two fold. From an inferential perspective, we wanted to understand how each of the measurements mentioned above (`Re`, `Fr`, `St`) affects the probability distribution for particle cluster volumes. From a predictive standpoint, we sought to build a model that would best predict the particle cluster volume distribution from observed values of each of these three parameters. In order to achieve the simultaneous objectives of interpretability and predictive performance, our group was careful in considering highly complex models. When building models for each of the four moments, we tried a handful of modelling methods (scaling in low to high complexity) on our training data, and produced both indirect and direct estimates of testing error from AIC and cross validation for each of these methods. We then made decisions on which model was most effective by considering both interpretability and our testing error estimates. Our reason for utilizing AIC and cross validation error as a measure of predictive performance was due to our lack of labeled testing data on which we could directly obtain a measure of testing error.

Methodology

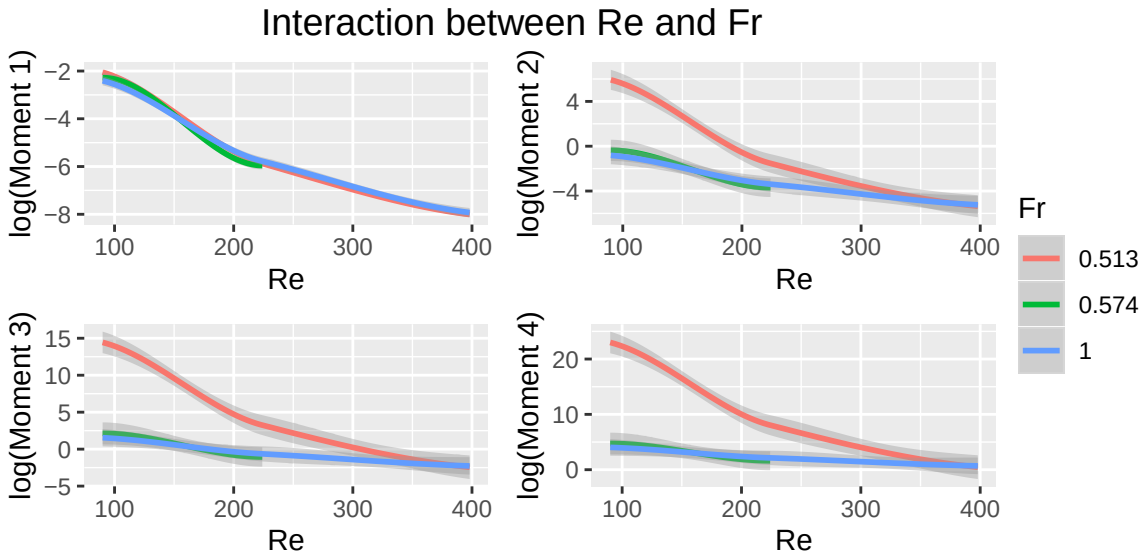
EDA

As mentioned in our introduction, in our attempt to model the probability distribution for particle cluster volumes, we decided on creating four separate models, one for each of the first four raw moments. However, prior to separating our models by moment as response variable, we wanted to explore general trends in our data that would inform any necessary transformations and/or interactions between the parameters. First, we knew that we needed to transform the Fr variable, as it contained infinity values. In order to do so, we used the inverse logit function ($\frac{e^x}{1+e^x}$), which transforms any real number x into a value in the interval [0,1].

Next, we visualized the distribution of the four moments. From this, we noticed a severe right skew in the distributions of all four moments. To combat this, we decided to do a log transformation of each of the moments. We can see that it results in more "normal" looking distributions.



Next, we wanted to explore potential interactions that would be influential within our models. Below we visualize the interaction between `Re` and `Fr` as it pertains to each moment.



As we can see above, there seems to be a notable interactive effect between `Re` and `Fr` for the second, third, and fourth moments. This does not seem to be the case in the first, where varied values of `Fr` do not yield markedly different slopes on the plot. Thus, in the cases of the second, third, and fourth moments, we will consider this interaction as we build models. We also considered the interaction effect between `St` and `Fr` and the interaction effect between `St` and `Re` (see appendix A), but there does not seem to be any notable patterns in these effects.

Something that is important to note is that, although our `Re` and `Fr` variables each take only one of three values, we are not treating them as categorical. If one of our goals is to be able to do prediction and extrapolate given any set of parameter values (perhaps `Re` and `Fr` values that are not strictly one of the three we see in the given training data), then it is important to consider these values as continuous.

Model Selection

When deciding on a final model for each of the moments, we tested a handful of models with varying complexities and interpretabilities. One constant was we used the log transformed moments as our reponse variable for each of our final models. For each moment, we tested various linear, polynomial, and GAM models. We created models with and without interactions, and then conducted ANOVA tests to determine that models with interactions are generally better (except for the first moment). We avoided testing out more complex models, such as random forest, for the reason that it would provide a very limited scope from an inferential standpoint.

For each of our final models, we also looked at the residual plots. These plots allowed us to observe linearity, constant variance, and normality of the residuals. In general, the distribution of residuals vs. the predictors and the QQ plot of the residuals seem decent. However, since there are limitations with our data, we have to be more lenient with these assumptions.

After creating various models for each moment, we made comparisons in both AIC and cross validation error. Our rationale for choosing 5 folds was the small size of the training dataset, which contained only 89 observations. Below is a summary of the performances of each model we tested, for each of the four moments.

Results

The final model output for each moment is displayed in the appendix, along with interpretation of predictors (see section B).

Moment 1

```
##                                CV.Error.Moment1 AIC.Moment1
## Log Linear with Log Predictors           0.0314       -54.6
## GAM (without interactions)               0.4110       171.0
## GAM (with interaction)                   0.3970       169.0
```

We are using the linear model with a logged response and logged `Re` and `Fr` predictors as our final model for moment 1. It has the lowest CV error and AIC. We initially did not log transform the predictors, but the residual plots for the untransformed predictors egregiously violate the assumptions; the residual plots for the transformed predictors are slightly better (see appendix C.1).

Moment 2

```
##                              CV.Error.Moment2 AIC.Moment2
## Log Linear with interactions             4.71         395
## Polynomial (degree 2)                    3.53          NA
## GAM (without interactions)               5.75         406
## GAM (with interaction)                   5.09         390
```

We are using the linear model as our final model for moment 2. Although it does not have the lowest error values, it is more interpretable than a polynomial. We considered log transforming `Re` and `Fr` like we did for moment 1, but the residual plots before and after transforming those predictors are noticeably different, so we kept the untransformed version of the predictors (see appendix C.2). In this linear model we are including an interaction between `Re` and `Fr`. To make this interaction effect tangible, we will take our second moment as an example. Fixing `Re` at 90, a unit increase in `Fr` will produce a $\beta_3 + 90(\beta_4)$, or -8.75 decrease on $\log(y)$, or a 1.0001 factor increase on y (where y is our second moment).

Moment 3

```
##                              CV.Error.Moment3 AIC.Moment3
## Log Linear                               15.7         498
## Polynomial (degree 2)                    12.1          NA
## GAM (without interactions)               16.3         512
## GAM (with interaction)                   13.3         497
```

We are using the GAM with interactions as our final model for moment 3. Even though the error is slightly lower for the polynomial model, it is more interpretable, especially when we are trying to consider the impact of multiple variables. We considered log transforming `Re` and `Fr` like we did for moment 1, but the residual plots before and after transforming those predictors are noticeably different, so we kept the untransformed version of the predictors (see appendix C.3).

Moment 4

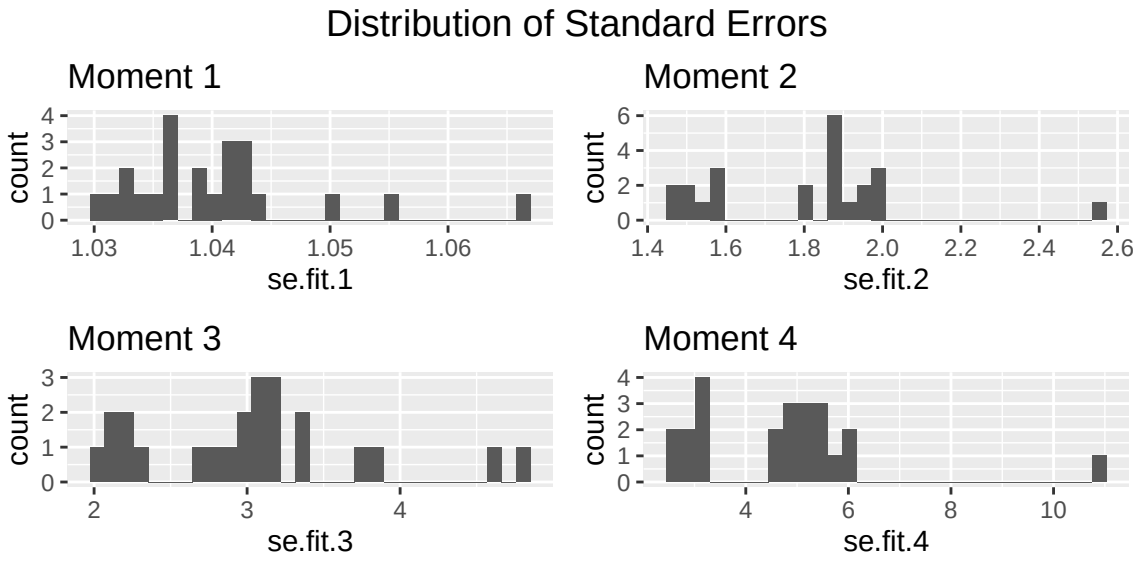
```
##                              CV.Error.Moment4 AIC.Moment4
## Log Linear                               30.6         562
## Polynomial (degree 2)                    23.1          NA
## GAM (without interactions)               38.8         578
## GAM (with interaction)                   32.1         562
```

We are using the GAM with interactions as our final model for moment 4. Even though the error is lower for polynomial model, it is more interpretable, especially when we are trying to consider the impact of multiple variables as we mentioned previously. We considered log transforming `Re` and `Fr` like we did for moment 1, but the residual plots before and after transforming those predictors are noticeably different, so we kept the untransformed version of the predictors (see appendix C.4).

Test Data Set

Measuring Uncertainty

In discussing the performance of our model on the testing data, it is important to acknowledge the uncertainties of our predicted values for each model. Below we visualize the distribution of the standard error values of our predictions for each of our moments. This will give us a visual understanding of the width of our confidence intervals for each prediction.



As we can see, for our predictions of the first moment, our standard errors for the predictions of the first moment are relatively small, highlighting a significant degree of certainty. As we shift focus to the uncertainty of prediction in the second moment, we see that the standard error of the fit is slightly larger, clustering at 1.5 and having its highest value of 2 (when St, Re, and Fr, take on values of 2, 90, and 1 respectively). Because this resulting confidence interval is going to be larger than other confidence intervals for this moment, the users of this model may be slightly more wary of the predicted values obtained. In terms of the third and fourth moments, we see that the standard error of fit hovers around even greater values from 3.5 to 6. Here, we are more uncertain of our predictions. Thus, for a user attempting to obtain predictions of the third and fourth raw moments from our models, it may be safer to seek out a direct value of the moment.

Conclusion

In conclusion, we found that, for each of the moments barring the first, there was a tradeoff between prediction and interpretability. It was often the case that the final model we chose was not the one with the lowest cross validation error or AIC value. However, in sacrificing estimated test error, we gained in interpretation: both linear models and GAMs have advantages over more complex models in their ability to demonstrate the effects of individual variables on each of the first four raw moments.

Across all four moments, judging from the p-values, all three parameters are individually important. We found it important to leave all variables in each of our models, as opposed to performing some sort of variable selection, as we were working with such little information as is. Furthermore, at higher moments, the interaction between `Re:Fr` becomes increasingly important. This is evidenced by the interaction plots in our exploratory data analysis, as well as our models with interactions (at higher moments) consistently having lower estimated testing errors.

However, despite our success in finding a model that met both our goals, it is important to acknowledge that there were some inherent limitations in the training data we worked with. These are limitations that will certainly have implications on how our final model will perform on the testing data. For example, though there was a lack of continuous values for `Re` and `Fr` within the data, we sought to build a model that would theoretically predict the four raw moments for any continuous values of `Re` and `Fr.` If our training data itself had more continuity in these values, perhaps predictive performance would be better.

In continuing with this expedition in the future, it would be interesting to delve more deeply into the correlations between the various moments. Other groups in class today discussed using chained equations to explore the relationships between the first four moments. Understanding these intricacies better could improve our modeling techniques and perhaps be of more use to the users in the future.