# Summary of "Statistical Data Analysis in the Computer Age"

February 1, 2017

## Introduction

The following document summarizes the work presented by Efron and Tibshirani titled "Statistical Data Analysis in the Computer Age" [1]. The paper presents a summary of four statistical techniques which require fewer underlying assumptions but more computationally intensive. The author argues that with the constant increase of computational power, these statistical test methods should be more common place as they provide a more accurate representation of data over classical statistics. The following review summarizes the four different techniques presented in the paper followed by a brief example of the application of a method.

## Bootstrapping

Bootstrapping is the extraction of any estimator from a given dataset. This is accomplished through sampling of a dataset of size n with replacement to produce another set also of size n for which the estimator can be recalculated. The mean estimator seen is a better reflection of the true population mean. The authors claim that this method will work for resampling as low as 25 time, although they recommend resampling greater than 200 as the standard error at this point would be negligible. This method does not require prior knowledge of the sample data. The sample dataset however needs to be representative of the population to provide any estimates about the population. This bootstrap technique is also not restricted by the data type of the individual points of the sample.

## Non-Parametric Regression and Generalized Additive Models

The second method discussed was non-parametric regression. Non-parametric regression involves loess fitting. Loess works by sectioning the data into windows for which local regression is performed for each. A tricube function, a smooth weight function, is then applied to the resulting regression

followed by a weighted linear regression for each window. The smooth weighted functions construct a generalized additive model for the data. Because this model is composed of a fit of multiple curves for a localized region, a perticular function is not imposed on the entire dataset which the author claims allows for better insight to the true nature of the dataset.

## Classification and Regression Trees (CART)

The final method discussed was the classification and regression tree (CART) method. This method based off of the form of a binary decision tree. The data is broken down in series at each node where a positive answers to a nodes assigned question are assigned to the left branch of a node and negative to the right. Final observation and class determination of the data are seen at the terminal/leaf nodes of the tree, where each node is assigned a class. Node classes are generally assigned through majority however weights may be added to increase cost of misclassifications for a particular class. The height of the final tree is determined through pruning whereby a large tree is constructed and pruned from the bottom to avoid overfitting. Ten different trees are generated through 10-fold cross validation (i.e. 90% of the data is used to train the tree and 10% used to test). The height of tree with the lowest misclassification rate is used as the height parameter for the final tree. The use of cross validation during training also has the added benefit of providing a better estimate of misclassification rates.

## Application

The methods discussed in this paper are provide methods of statistical testing which require fewer underlying assumptions which results in a more accurate statistical test. One problem I am trying to solve is the prediction of protein interaction given a landscape of potential interaction regions. The method of CART could be useful in the construction of a decision tree to aid in the automatic classification of protein interaction. The method of bootstrapping the data could be used during CART classification to further improve estimation.

## References

[1] B. Efron and R. Tibshirani, "Statistical data analysis in the computer age," *Science (80-. ).*, vol. 253, no. 5018, pp. 390–395, 1991.