Form Fitting Random Variable Distribution Estimation

M.C. Rumuly
ECEN 662-600
Texas A&M University
College Station, TX

***Abstract*—Objective functions were evaluated for the purpose of choosing a general distribution to represent a random system given a sample of a certain size. As they converged in accuracy and represented intuitively closer distributions with statistic values indicating closer distributions, the statistics were considered successful. The Cramér-von Mises Criterion had the maxi-min performance across all real and assumed distributions, with one of the least-complex calculations, making it the most versatile 'goodness of fit' test considered.**

I. Introduction

In constructing models to analyze unknown random systems, it is important to be able to choose the most appropriate distribution to represent the underlying randomness. Much rigor has been spent on the problem of estimating the parameters of a distribution where the basic underlying form is assumed or known in advance. However, answering the estimation question raises the issue of choosing an appropriate basic model in the first place. One method is to construct a histogram from the sample and use intuition about the shape to infer a basic distribution to estimate; this project seeks to explore a more rigorous approach.

This project seeks to find a general 'goodness of fit' criterion which answers the following question: Given a sample of independent, identically distributed (IID) observations $\boldsymbol{X} = \{x_1, x_2, \dots, x_n\}$ generated by a random variable (RV) with continuous distribution $F(\cdot)$, which of a pre-defined set of generic distributions $\widehat{\boldsymbol{F}} = \{\widehat{F_1}, \widehat{F_2}, \dots, \widehat{F_m}\}$ best approximates the real distribution? This question is considered in a classical framework of statistics, without Bayesian priors or costs associated with the different distributions, in order to focus exclusively on the comparison between distributions.

Fundamentally, the sample $\boldsymbol{X}$ can be represented as a distribution function itself, with its own cumulative distribution function (CDF) known as the Empirical Distribution Function (EDF), which has the form of (1), where $\mathbf{1}_A$ is the indicator function of event A [1]. Notice that the derivative of this function, the empirical equivalent of a probability density function (PDF), is the series of delta functions in (2).

$$F_{\boldsymbol{X}}(t) = \frac{1}{n}\sum_{i=1}^{N} \mathbf{1}_{x_i \le t} \tag{1}$$

$$f_{\boldsymbol{X}}(t) = \frac{1}{n}\sum_{i=1}^{N} \delta(t - x_i) \tag{2}$$

Furthermore, the Glivenko-Cantelli Theorem states that as the size of the sample from a given distribution increases, the EDF converges to the generating distribution almost surely [2].

$$\lim_{n \to \infty} \left(\sup_{t \in \mathbb{R}} |F_{\boldsymbol{X}}(t) = F(t)| \right) \to 0 \tag{3}$$

While the EDF is an increasingly accurate representation of the underlying distribution, it has many undesirable characteristics in analysis for finite n: most importantly, it is non-smooth and not capture probability densities for unobserved values, even in close proximity to observed values. Thus it is important to find the best $\widehat{\boldsymbol{F}}$.

II. Objective Functions

A set of objective functions are compared to evaluate their behavior. The properties desired of each statistic are: strictly greater-than or equal to 0, such that a greater value indicates a worse fit; calculability for non-smooth distributions to allow for the EDF; values comparable across all generic distributions without correction for critical values; low computation complexity. The following four statistics were found to satisfy the constraints and were evaluated.

As an aside, each statistic is capable of rejecting an individual distribution based on critical values associated with that distribution, this capability is not required to choose the best approximation from a finite set. Furthermore, the dictionary of critical values necessary to utilize this capability is rendered invalid by the use of estimated parameters. As such, this application is not implemented or discussed further in this paper

A. Kolmogorov-Smirnov Test [3]

This simple test uses the maximum absolute error between two CDFs, or an EDF and CDF. The statistic generated when comparing the EDF of a sample and the CDF of a distribution is (4).

$$S_{KS}(\boldsymbol{X}, \hat{F}_k) = \sup_{\mathrm{t}} \left| F_{\boldsymbol{X}}(t) - \hat{F}_k(t) \right| \tag{4}$$

B. Kuiper's Test [4]

This test is a refinement on the Kolmogorov-Smirnov Test, which creates a statistic equally sensitive to the tails and median of distributions, and is invariant under cyclic transformations of the independent variable. It does so by taking into account the greatest negative and positive differences between the two distributions being compared.

$$S_{KT}(\boldsymbol{X}, \hat{F}_k) = \max_{\mathrm{i}} \left[\frac{i}{n} - \hat{F}_k(x_i) \right] + \max_{\mathrm{i}} \left[\hat{F}_k(x_i) - \frac{i-1}{n} \right] \tag{5}$$

C. Anderson-Darling Test [5]

This test is based on the mean square error between the EDF and CDF being compared, being the most general form of the quadratic class of EDF statistics. It is based on the weighted squared distance between two distributions in (6).

$$A^2(\boldsymbol{X}, \hat{F}_k) = n \int_{-\infty}^{\infty} \frac{\left(F_{\boldsymbol{X}}(t) - \hat{F}_k(t) \right)^2}{\hat{F}_k(t) \left(1 - \hat{F}_k(t) \right)} dF(x) \tag{6}$$

The basic, most direct test statistic from this distance can be represented as a Riemann Sum. This statistic places most of the weight on the tails of the distribution. Having ordered the sample $\boldsymbol{X}$ from least to greatest to produce $\boldsymbol{Y} = \{y_1 < y_2 < \cdots < y_n\}$, the statistic is (7). The use of logarithms makes this statistic hard to calculate where $\hat{F}_k(y_i)$ is arbitrarily close to 0 or 1. In the implementation for this project, any y_i which produces an undefined value is excluded from the statistic.

$$S_{AD}(\boldsymbol{Y}, \hat{F}_k) = -n - \sum_{i=1}^{n} \left(\frac{2i-1}{n} \left[\ln\left(\hat{F}_k(y_i) + \ln\left(1 - \hat{F}_k(y_{n+1-i}) \right) \right) \right] \right) \tag{7}$$

D. Cramér-von Mises Criterion [6]

This test is a simplified version of Anderson-Darling which uses an alternate weight on the area function, eliminating the need for logarithms. This simplifies the calculation and removes the problem of undefined values. Using the same ordered set $\boldsymbol{Y}$, the statistic is (8).

$$S_{CM}(\boldsymbol{Y}, \hat{F}_k) = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - \hat{F}_k(y_i) \right)^2 \tag{8}$$

III. Distributions

In order to analyze the behavior of the objective functions, samples were generated from and then compared against a set of four distributions.

Chosen a set of distributions with which to produce and characterize random samples. Note: parameters (to be specified/estimated, estimation used, default values for generation), pdf, cdf

A. Uniform Continuous

This is perhaps the most basic ubiquitous probability distribution. The two parameters, a and b, are the lower and upper bounds of the valid interval, where $a < b$, and the default values are $a = 0, b = 1$. For an unbiased estimator of both parameters, the midpoint and the maximum absolute distance from the midpoint were estimated in (11) and (12). The maximum-likelihood estimator was avoided, as it always forces at least two observations to produce undefined Anderson-Darling values.

$$f_U(t; a, b) = \begin{cases} \dfrac{1}{b-a}, & a \le t \le b \\ 0, & otherwise \end{cases} \tag{9}$$

$$F_U(t; a, b) = \begin{cases} 0, & t < a \\ \dfrac{t-a}{b-a}, & a \le t \le b \\ 1, & t > b \end{cases} \tag{10}$$

$$\hat{a}(\boldsymbol{X}) = \frac{\frac{1}{n}\sum \boldsymbol{X} + median(\boldsymbol{X})}{2} - \frac{n+1}{n}\max_{X}\left|x_i - \frac{\frac{1}{n}\sum \boldsymbol{X} + median(\boldsymbol{X})}{2}\right| \tag{11}$$

$$\hat{b}(\boldsymbol{X}) = \frac{\frac{1}{n}\sum \boldsymbol{X} + median(\boldsymbol{X})}{2} + \frac{n+1}{n}\max_{X}\left|x_i - \frac{\frac{1}{n}\sum \boldsymbol{X} + median(\boldsymbol{X})}{2}\right| \tag{12}$$

B. Normal

Between the usefulness of its additive properties and in analyzing white noise, often called white Gaussian noise, the normal distribution is a deeply important and widely used distribution. The two parameters, μ and σ^2, are the mean and variance of the distribution with default values $\mu = 0$ and $\sigma^2 = 1$. The variance estimation uses the maximum-likelihood estimator (16).

$$f_N(t;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} \tag{13}$$

$$F_N(t;\mu,\sigma^2) = \Phi\left(\frac{t-\mu}{\sqrt{\sigma^2}}\right) \tag{14}$$

$$\hat{\mu}(\boldsymbol{X}) = \frac{1}{n}\sum \boldsymbol{X} \tag{15}$$

$$\widehat{\sigma^2}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{16}$$

C. Exponential

Unlike the uniform distribution, which is confined to a finite interval, and the normal distribution, which extends infinitely in both the positive and negative directions, the exponential distribution has a lower bound, but extends positively to infinity. It is useful for modeling arrival times in certain systems, and has a useful memoryless property. Its two parameters, l and β, denote the lower bound and scale of the distribution, and are assigned default values of $l = 0$ and $\beta = 1$. The $\beta = \frac{1}{\lambda}$ scale parameter was chosen for the existence of a simple unbiased estimator, in order that an unbiased estimator of l could be used (19). The maximum-likelihood estimator for l was avoided, as it always forces the minimum observation to produce an undefined Anderson-Darling value. The unbiased β estimator in (20) makes use of the memoryless property to avoid circular estimation.

$$f_E(t;l,\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{t-l}{\beta}}, & t \geq l \\ 0, & t < l \end{cases} \tag{17}$$

$$F_E(t;l,\beta) = \begin{cases} 1 - e^{-\frac{t-l}{\beta}}, & t \geq l \\ 0, & t < l \end{cases} \tag{18}$$

$$\hat{l}(\boldsymbol{X}) = \min \boldsymbol{X} - \frac{1}{n^2}\sum \boldsymbol{X} \tag{19}$$

$$\hat{\beta}(\boldsymbol{X}) = \frac{1}{n}\sum \boldsymbol{X} - \min \boldsymbol{X} \tag{20}$$

D. Laplace

Perhaps the least frequently used of the distributions included here, the Laplace distribution is most useful for modeling the difference between two IID exponential arrival times. It was included here for its similarity to the normal distribution in range and symmetry, and similarity to the exponential distribution in shape, allowing for a more challenging element by which to evaluate the objective functions. It contains two parameters, a location μ and a scale b, which are given the default values $\mu = 0$ and $\beta = 1$.

$$f_L(t;\mu,b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \tag{21}$$

$$F_L(t;\mu,b) = \begin{cases} \frac{1}{2} e^{\left(\frac{x-\mu}{b}\right)}, & x \leq \mu \\ 1 - \frac{1}{2} e^{\left(-\frac{x-\mu}{b}\right)}, & x > \mu \end{cases} \tag{22}$$

$$\hat{\mu}(\boldsymbol{X}) = median(\boldsymbol{X}) \tag{23}$$

$$\hat{b}(\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{\mu}| \tag{24}$$

IV. Test Method

A simulation was used to explore the behavior and accuracy of the objective functions with respect to each distribution. This simulation was implemented in Python 3.6; the source code can be found in the associated GitHub project folder at https://github.com/CourseReps/ECEN662-Spring2018/tree/master/Students/mason-rumuly/FinalProject. This implementation uses object oriented and functional programming techniques to be easily scalable, where new distributions and objective functions can be easily added.

For a given distribution $F_i(t)$ with the default parameters, generate a random sample $\boldsymbol{X} = \{x_1, x_2, \dots, x_n\}$ of size n. Now, for each distribution $F_j(t)$, estimate the parameters of that distribution assuming the sample came from that distribution to get $\hat{F}_j(t|\boldsymbol{X})$. Compute the statistic of each objective function. Record the value of each objective function for this ordered pair $S_o(i,j)$; this statistic reflects the appropriateness of the assumption that $i = j$. Finally, for each objective function, record $A_o(i) = 1_{\min_j S_o(i,j) \forall i=j}$ as the accuracy of objective function o.

This simulation was run for $n = \{8, 10, 20, 40, 80, 100, 200, 400, 800, 1000\}$. All values were averaged over 1000 runs of the simulation in order to provide a more robust data set.

V. Results

The simulation was successfully run for all distributions and recorded for all objective functions. The convergence to 100% in accuracy is shown in Fig. 1 through Fig. 4. The behavior of each objective function with each generator-guess pair is shown in Fig. 5 through 20.

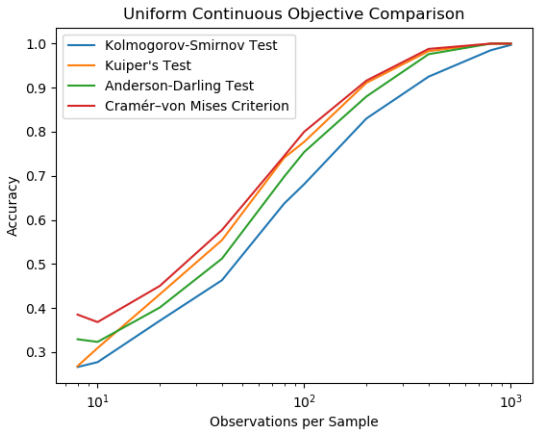

Figure 1: Average Uniform Objective Accuracy

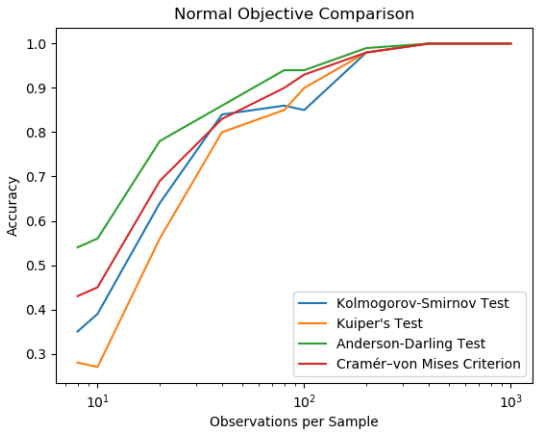

Figure 2: Average Normal Objective Accuracy

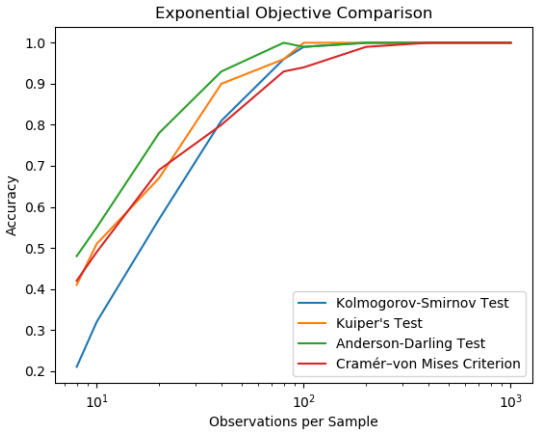

Figure 3: Average Exponential Objective Accuracy

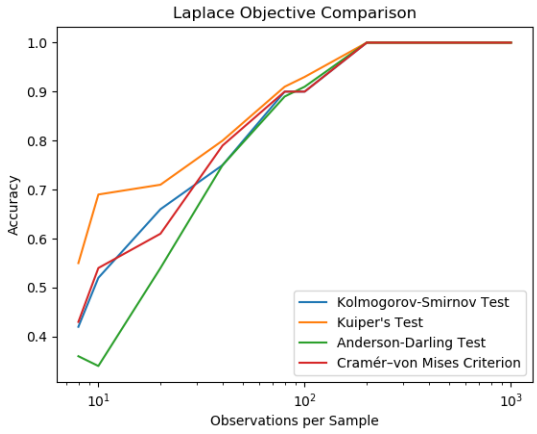

Figure 4: Average Laplace Objective Accuracy

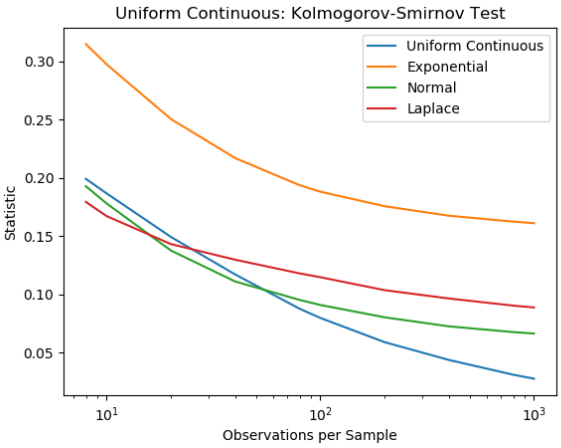

Figure 5: Average KS Behavior Uniform

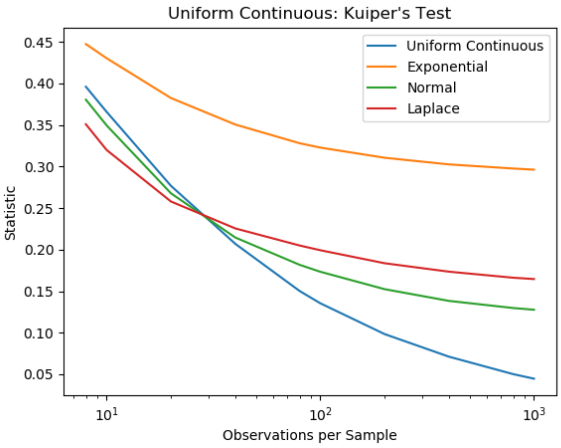

Figure 6: Average KT Behavior Uniform

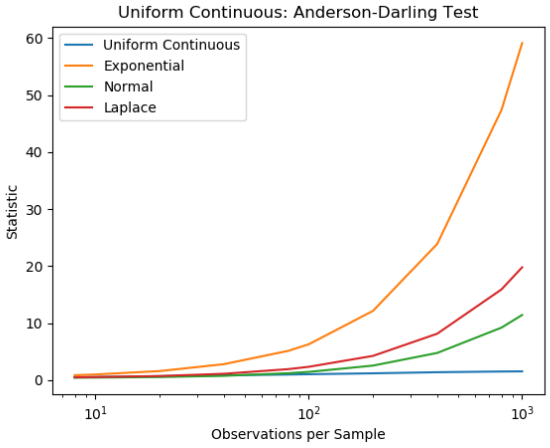

Figure 7: Average AD Behavior Uniform

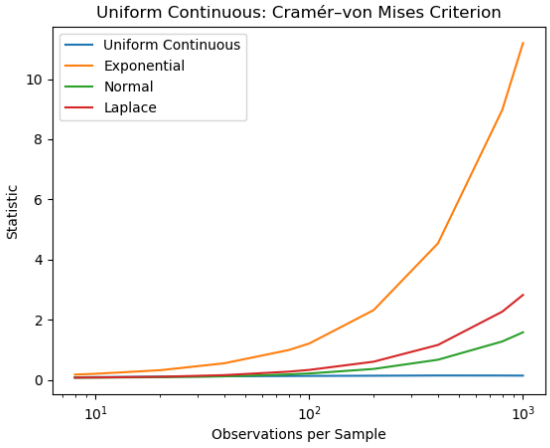

Figure 8: Average CM Behavior Uniform

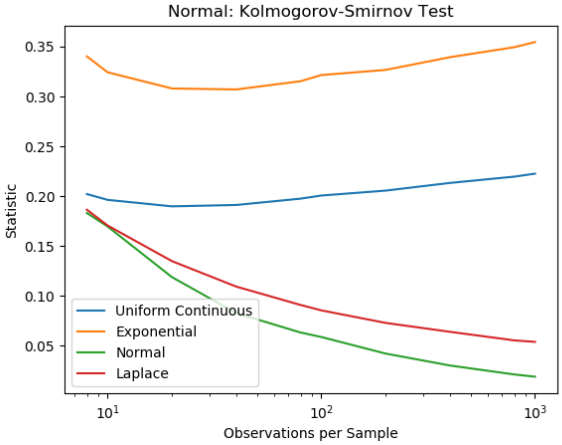

Figure 9: Average KS Behavior Normal

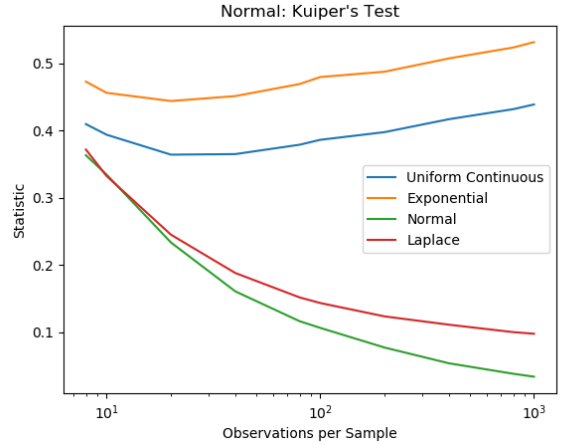

Figure 10: Average KT Behavior Normal

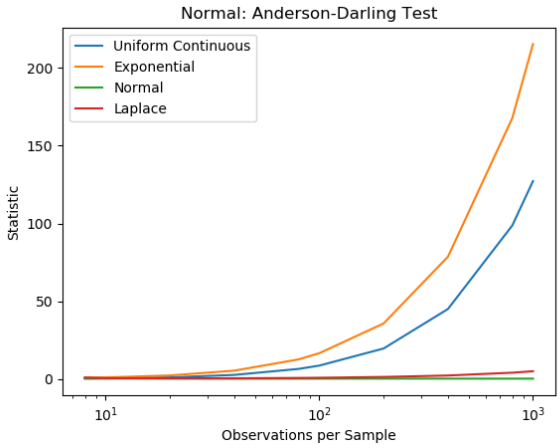

Figure 11: Average AD Behavior Normal

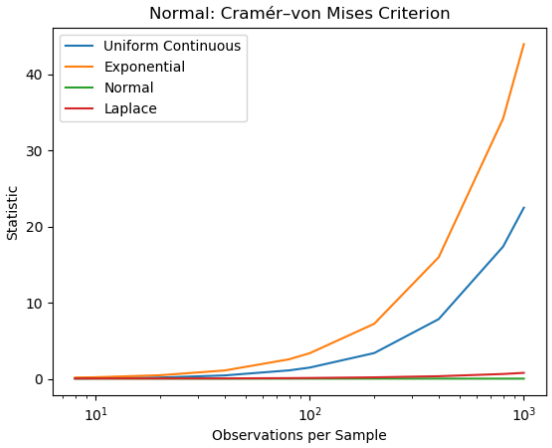

Figure 12: Average CM Behavior Normal

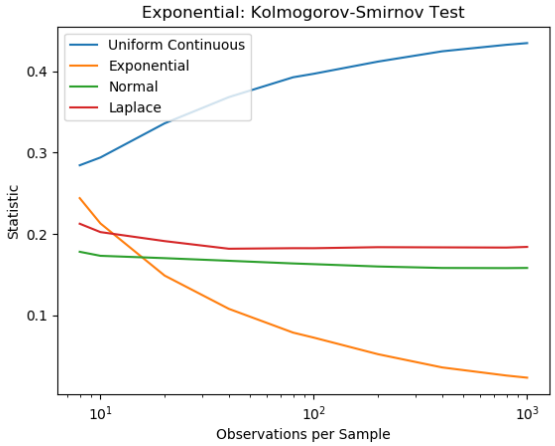

Figure 13: Average KS Behavior Exponential

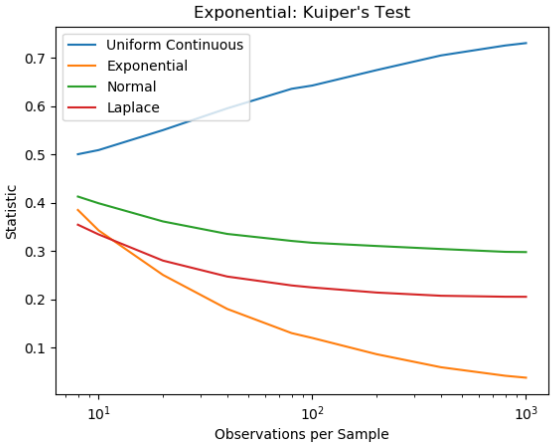

Figure 14: Average KT Behavior Exponential

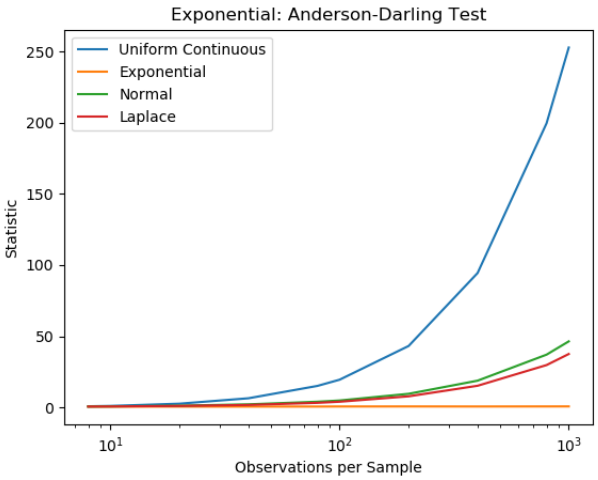

Figure 15: Average AD Behavior Exponential

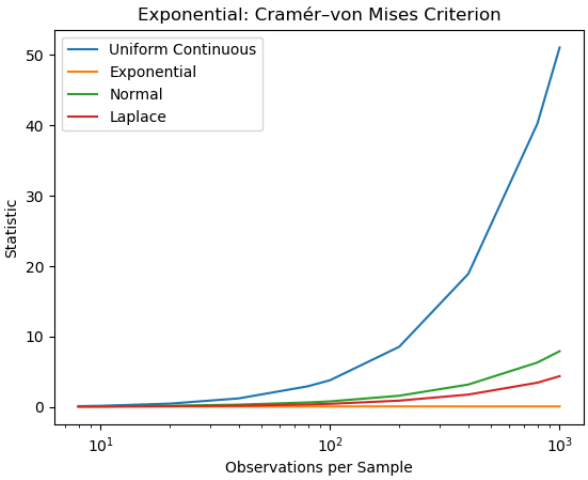

Figure 16: Average CM Behavior Exponential

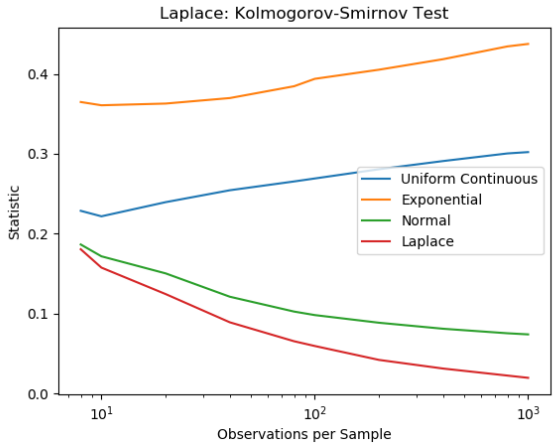

Figure 17: Average KS Behavior Laplace

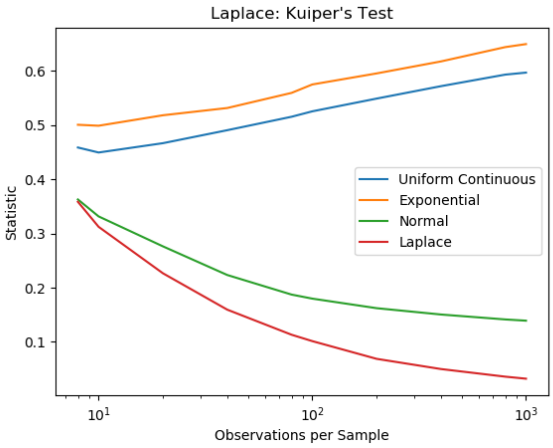

Figure 18: Average KT Behavior Laplace

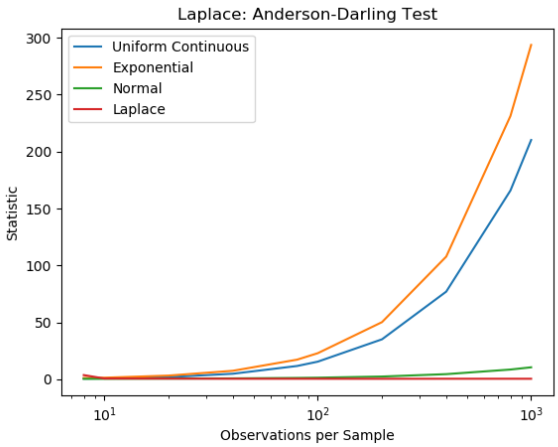

Figure 19: Average AD Behavior Laplace

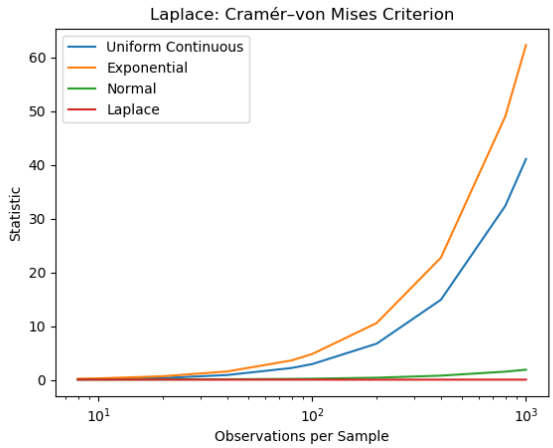

Figure 20: Average CM Behavior Laplace

VI. Conclusions

The convergence of each objective statistic to 100% accuracy shows that they correctly identify the important features of each distribution. Note also from the generator-guess pair behavior that the distribution which is intuitively more similar tends to be a distant second-best, showing that it also identifies similarities when the exact distribution is not part of the allowed set, e.g. if the Laplace were excluded from guessing, on average they identified the Normal distribution where the Laplace is the real distribution.

Of all the objective functions, the Cramér-von Mises Criterion demonstrated the most versatility. Although it is only the most accurate function for the uniform distribution, it has a middle-of-the-pack accuracy for all other distributions in Fig. 2-4. It is also an easy function to calculate. Since the Kolmogorov Smirnov Test and Kuiper Test statistics are approximately the same complexity with less average accuracy, the Cramér-von Mises Criterion is strictly better.

The Anderson-Darling Test statistic was the most accurate discriminator for two of the four distributions, but due to its inability to handle outliers with its logarithms is more unwieldy to use, and is the worst discriminator for the Laplace distribution.

Overall, the goals of this project were accomplished. This work could be continued by adding more distributions, analyzing the convergence speed by samples, or by creating a Bayesian framework which incorporates priors into the statistics, among other things.

References

[1] van der Vaart, A.W. (1998). Asymptotic statistics. Cambridge University Press. p. 265. ISBN 0-521-78450-6.

[2] Howard G.Tucker (1959). "A Generalization of the Glivenko-Cantelli Theorem". The Annals of Mathematical Statistics. 30: 828–830. doi:10.1214/aoms/1177706212.

[3] *Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons". Journal of the American Statistical Association. American Statistical Association. 69 (347): 730–737. doi:10.2307/2286009. JSTOR 2286009.*

[4] *Pearson, E.S., Hartley, H.O. (1972) Biometrika Tables for Statisticians, Volume 2, CUP. ISBN 0-521-06937-8 (page 118)*

[5] *Anderson, T.W.; Darling, D.A. (1954). "A Test of Goodness-of-Fit". Journal of the American Statistical Association. 49: 765–769. doi:10.2307/2281537.*

[6] *Anderson, T. W. (1962). "On the Distribution of the Two-Sample Cramer–von Mises Criterion" (PDF). Annals of Mathematical Statistics. Institute of Mathematical Statistics. 33 (3): 1148–1159. doi:10.1214/aoms/1177704477. ISSN 0003-4851.*