Effective Nonparametric Distribution Modeling for Distribution Approximation Applications

Thomas C. H. Lux[1], Layne T. Watson[2,5], Tyler H. Chang[1], Li Xu[3], Yueyao Wang[3], Jon Bernard[1], Yili Hong[4], Kirk W. Cameron[2]

***Abstract*—Many fields of science rely on the collection of samples and estimation of true population distributions from those samples. There are several effective nonparametric methods for approximating a true distribution from empirical data, however it is unclear which methods produce the best approximations in practice. This work presents a case study on the effectiveness of various distribution approximations. Results show that piecewise linear approximations produce the smallest maximum absolute error, while the classic empirical distribution function (EDF) produces the smallest median absolute error as well as the smallest first quartile and minimum absolute error when approximating a distribution from a sample. When building distribution prediction models, the piecewise quintic and cubic approximations produce the lowest absolute error at most error percentiles. This case study encourages more research on the best methods of fitting empirical data with smooth functions to generate accurate distribution approximations.**

I. INTRODUCTION

Empirical samples play a pivotal role in science. Experiments are run, data is recorded, and that data is used to draw conclusions about the *truth*. When an experiment is run many times with varying outcomes, it is common to describe the truth as a random variable. In this work continuous (numeric) outcome random variables are considered. In this context, a random variable X is precisely defined by its absolutely continuous *cumulative distribution function* (CDF) F_X and the derivative of the CDF, the *probability density function* (PDF) f_X. As samples are drawn from X, the value of the CDF can be estimated at the sample points by measuring the probability at which other samples are less than or equal to that value. In this sense empirical data defines an empirical distribution function (EDF). The points that define the EDF are used to construct approximations of the true underlying CDF.

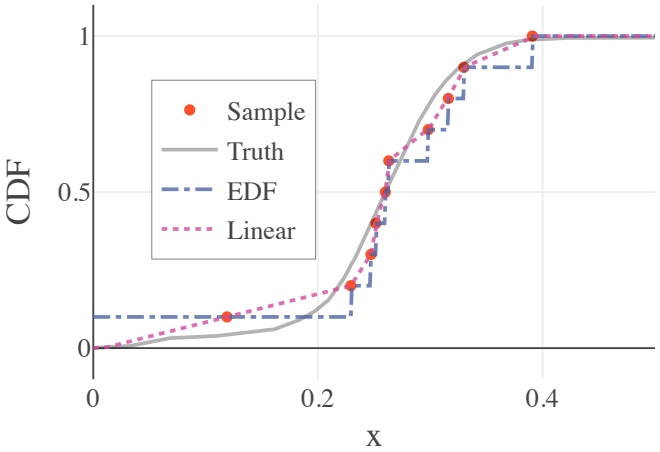

Fig. 1. A demonstration of the linear fit and the classic EDF over a sample of 10 points. This data is normalized to be in the range $[0, 1]$ and the right half of the domain is cropped out of this plot. Notice that the EDF points do not necessarily (and often do not) exactly equal the true CDF at any given position. The sampling inherently introduces variance into measurements of the CDF. This makes producing an accurate approximation difficult, as will be seen.

An approximation of a CDF can have varying levels of smoothness and, for empirical purposes the approximations are constructed over a closed interval. Piecewise polynomial functions (splines) provide the smooth approximations for animation in graphics [1], [2], aesthetic structural support in architecture [3], efficient aerodynamic surfaces in automotive and aerospace engineering [3], and most importantly to this work they can provide accurate nonparametric approximations in statistics [4]. While polynomial interpolants or regressors apply broadly, splines are often a good choice because they can approximate globally complex functions while minimizing the local complexity of an approximation.

In this statistical work, the construction of a monotone interpolating spline that is continuous in its derivatives could be meaningfully useful [5]. A function with C^1, and especially C^2 continuity could approximate a cumulative distribution function to a high level of accuracy with relatively few intervals. A twice continuously differentiable approximation to a cumulative distribution function (CDF) would also pro-

*This work was supported by the National Science Foundation Grants CNS-1565314 and CNS-1838271.

[1]Doctoral candidate, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24060, USA (tchlux at vt.edu)

[2]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24060, USA

[3]Doctoral candidate, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24060, USA

[4]Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24060, USA

[5]Departments of Mathematics and Aerospace Engineering, Virginia Polytechnic Institute and State University, Virginia 24060, USA

978-1-7281-6861-6/20/$31.00

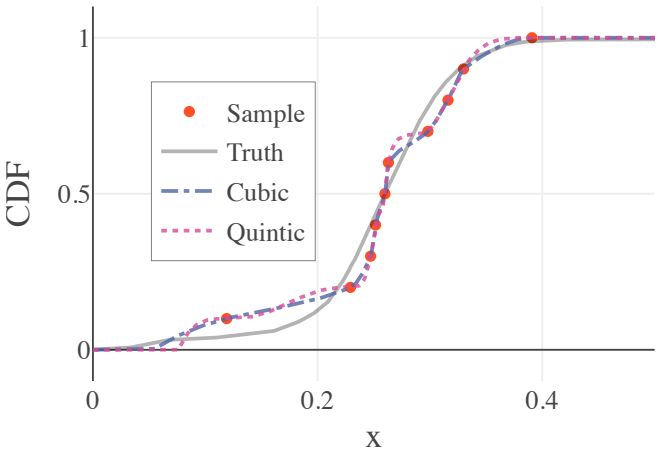

Fig. 2. A demonstration of the monotone cubic and quintic spline interpolants over a sample of 10 points. This data is normalized to be in the range $[0, 1]$ and the right half of the domain is cropped out of this plot. Notice that the difference between EDF points and the true CDF reduce the accuracy of first derivative (for both) and second derivative (for quintic) estimates in the interpolants, magnifying the error in the empirical estimations of CDF values.

duce a corresponding probability density function (PDF) that is continuously differentiable, which is a desirable property many standard parametric distributions maintain. In effect, an improved distribution estimate can increase the accuracy of distribution predicting models [6].

There is publicly available software for monotone piecewise polynomial interpolation, including quadratic [7], cubic [8], and (with limited application) quartic [9], [10], [11] cases. Theory has been provided for the quintic case [12], [13]. Recently an algorithm for the construction of monotone quintic splines has been produced as well [14]. This work considers the EDF, a linear interpolant, a monotone cubic spline interpolant, and a monotone quintic spline interpolant as candidate approximations.

In the next section, the methodology of this work is outlined and experimental setup is detailed. In Section III, three experiments related to the approximation of cumulative distribution functions are presented and analyzed. Finally Section IV concludes.

II. METHODOLOGY AND DATA

In order to identify the best performing approximation techniques for distribution estimation, a case study on real-world data is presented and three experiments are run on that data to test various fits. First the accuracy of distribution approximations with varying sample sizes is studied, then the likelihood of any technique having the smallest maximum error is estimated, and finally the approximations are used to make predictions in a distribution prediction application similar to [15]. Each experiment presents a unique, albeit with limited data, perspective on the quality of approximation provided by the four distribution approximation techniques.

The CDF modeling case study is constructed from a four-dimensional dataset produced by executing the IOzone benchmark [16] on a homogeneous cluster of computers. Each node

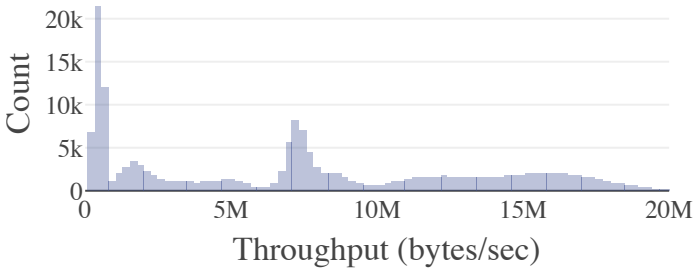

Fig. 3. A histogram of all throughput values across the 17 unique system configurations and 170 thousand executions of a "readers" test in IOzone.

contains two Intel Xeon E5-2637 CPUs offering a total of 16 CPU cores with 16GB of DRAM. While the CPU frequency varies depending on the test configuration, the I/O from IOzone is performed by an ext4 filesystem sitting above an Intel SSDSC2BA20 SSD drive. At the time of data collection, Linux kernel Version 4.13.0 was used. The system performance data was collected over two months by executing IOzone 10000 times for each of a select set of 17 system configurations, for a total of 170 thousand executions of IOzone. A single IOzone execution reports the max I/O throughput in bytes per second seen for the selected test type. For this case study, only the results of a "readers" test are considered, where bytes are sequentially read from the SSD. The summary of the data for the experiments for this paper can be seen in Table I. Distributions of raw throughput values being modeled can be seen in Figure 3.

The performance of approximation techniques that predict probability functions can be analyzed through a variety of summary statistics. The first two experiments study the distribution of absolute differences between approximated CDFs and the true CDFs. This distribution over many trials gives an idea of the *expected* error. The last experiment in this work analyzes the max absolute difference between approximated and true CDFs, also known as the Kolmogorov-Smirnov (KS) statistic [17] for its compatibility with the KS test.

The two-sample KS test is a useful nonparametric test for comparing two CDFs while only assuming stationarity, finite mean, and finite variance. The null hypothesis (that two CDFs come from the same underlying distribution) is rejected at level $p \in [0, 1]$ when

$$KS > \sqrt{-\frac{1}{2}\ln\left(\frac{p}{2}\right)}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

with distribution sample sizes $n_1, n_2 \in \mathcal{N}$. For all applications of the KS test presented in this work $n_1 = n_2$.

Finally an example of the round-trip prediction methodology from known and predicted distributions to the calculation of error can be seen in Figure 4. The Delaunay triangulation is a well-studied geometric technique for producing an interpolant [18]. The Delaunay triangulation of a set of data points into simplices is such that there are no data points inside the sphere defined by the vertices of each simplex. For a d-simplex

System Parameter	Values
File Size (KB)	4, 64, 256, 1024, 8192, 16384, 32768, 65536
Record Size (KB)	4, 8, 16, 32, 64, 128, 4096, 8192, 16384
Thread Count	8, 16, 24, 32, 40, 48, 56, 64
Frequency (GHz)	1.2, 1.6, 2, 2.3, 2.8, 3.2, 3.5

TABLE I
A DESCRIPTION OF SYSTEM PARAMETERS CONSIDERED FOR IOZONE. RECORD SIZE MUST BE LESS THAN OR EQUAL TO FILE SIZE DURING EXECUTION. IN ALL, 10 THOUSAND REPEATED TRIALS ARE RUN AT 17 UNIQUE SYSTEM CONFIGURATIONS.

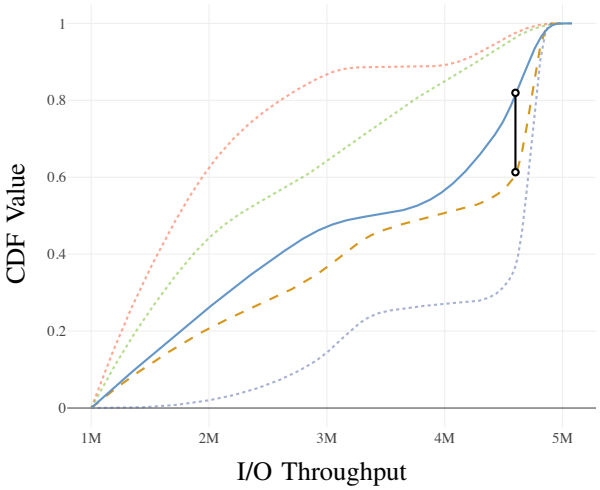

Fig. 4. In this example, the general methodology for predicting a CDF and evaluating error can be seen. The Delaunay method chose three source distributions (dotted lines) and assigned weights {.1, .3, .6} (top to bottom at middle). The weighted sum of the three known CDFs produces the predicted CDF (dashed line). The KS Statistic (vertical line) computed between the true CDF (solid line) and predicted CDF (dashed line) is 0.2 for this example. The KS test null hypothesis is rejected by p-value 0.01, however it is not rejected by p-value 0.001.

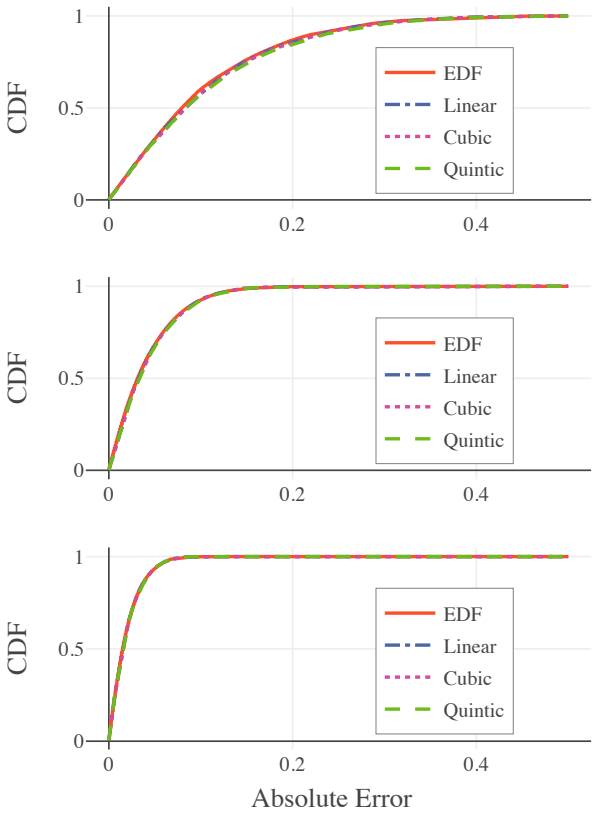

Fig. 5. The distribution of absolute errors with 10 (top), 50 (middle), and 200 (bottom) samples for each approximation method. Notice that the EDF performs slightly better than other methods (more small absolute errors) when there are only 10 samples. Given 50 or more samples, all the approximations produced nearly identical errors.

S with vertices $x^{(0)}$, $x^{(1)}$, ..., $x^{(d)}$, and functions $F_{x^{(i)}}$, $i = 0$, ..., d, $y \in S$ is a unique convex combination of the vertices,

$$y = \sum_{i=0}^{d} w_i x^{(i)}, \quad \sum_{i=0}^{d} w_i = 1, \quad w_i \geq 0, \quad i = 0, \ldots, d,$$

and the Delaunay interpolant F_y at y is given by

$$F_y = \sum_{i=0}^{d} w_i F_{x^{(i)}}.$$

In the case of these experiments, the file size, record size, thread count, and CPU frequency are first normalized to the unit hypercube. After normalization, Delaunay predictions are made via a leave-one-out method resulting in throughput data from five known configurations being used to predict one unknown throughput distribution. The Delaunay method is used to select the five source distributions as well as provide the convex weights associated with each of the source distributions that will most likely produce the predicted distribution.

III. RESULTS

All experiments that follow will consider the differences between CDF approximations given by the EDF, a linear interpolant, a monotone cubic interpolant, and a monotone quintic interpolant. In the first experiment, all four methods are used to approximate the "true" distributions of the ten thousand throughput values at each system configuration. One hundred random collections of k samples are drawn from each true distribution for $k = 10, 50, 200$ and all methods are used to approximate the true distribution from each sample. In Figure 5, the distributions of absolute difference between the "true" CDF and each approximated CDF at 1000 equally spaced percentiles is shown. The sample of 50 produces almost exactly half the errors observed with 10 samples, and the error is roughly halved again when increasing to 200 samples. Aside from the reduction in accuracy caused by sample size, most approximations appear to be nearly identical. When only 10 samples are observed, the EDF has slightly less absolute error than other techniques between the median and third quartile. This difference is very minor, and difficult to observe.

The second experiment considers the KS statistic (maximum absolute error) rather than the aggregate of errors. An increased collection of sample sizes is considered, with $k = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300, 400, 500$. For all k, 100 random samples of k throughput values are drawn from the true distributions and approximate CDFs are constructed via each of the four methods. The error of each approximate CDF is measured at 1000 equally spaced percentiles and the maximum absolute error is recorded. At each

sample size, Figure 6 depicts the probability that any given approximation method produces a distribution estimate with the lowest KS statistic. In this measure of error, the piecewise quintic and cubic approximations to the CDF provide the most accurate estimate of the true CDF. This result reveals that the chosen measure of error is important in determining which CDF approximation method is most suited to an applications. The EDF provides the lowest expected absolute error, however, the more smooth approximations appear to provide lower maximum absolute error.

The third experiment considers a distribution prediction application. In this case, the Delaunay method was applied to choose five source distributions and convex weights that will predict the throughput distribution at the system configuration with a 2.8GHz CPU frequency, 16 megabyte (MB) file size, 16MB record size, and 64 threads. This configuration is predicted because it is the only interpolation point (configuration inside or on the convex hull of other configurations) among the 17 available system configurations. An example of the prediction methodology can be seen in Figure 4. The distribution of absolute errors measured at 1000 equally spaced percentiles of the true CDF for all four approximation methods can be seen in Figure **??**. In this case the piecewise quintic and cubic approximations provide the best predictions overall. The quintic and cubic methods produce smaller errors than the piecewise linear and EDF approximations at more than 90% of measurements.

The three experiments presented in this section have each tested a unique facet of distribution approximation. The first experiment analyzes the approximation of a distribution from a sample when measuring the aggregate absolute CDF approximation error; in this case the EDF largely produces the best approximations. The second experiment analyzes the maximum absolute error when approximating a CDF from a sample; in this case the monotone quintic and cubic spline interpolants are most likely to provide the best approximation. The third and final experiment analyzes the aggregate absolute error when predicting unobserved distributions, for which the monotone quintic and cubic splines produced the best overall CDF approximations.

IV. CONCLUSION

This preliminary case study on the best functions for approximating cumulative distribution functions from empirical data has given some actionable insights. Specifically, the choice of error measure is important in deciding which distribution approximation method to apply to a sample. The standard empirical distribution function has the lowest expected absolute error when approximating a CDF given a sample of data. However, more smooth estimates like monotone quintic splines produce lower expected KS statistics (maximum absolute error), and are expected to have lower error in distribution prediction applications.

Future research may extend this case study to other data sets and include a more exhaustive distribution prediction test suite (experiment three). Tangentially, future work may target improved estimates of the first and second derivatives of the true CDF given a sample of data in order to reduce the approximation error of the monotone quintic and cubic spline methods.

REFERENCES

[1] D. L. Herman and M. J. Oftedal, "Techniques and workflows for computer graphics animation system," December 2015, uS Patent 9,216,351.

[2] A. Quint, "Scalable vector graphics," *IEEE MultiMedia*, vol. 10, no. 3, pp. 99–102, July 2003.

[3] A. Brennan, "Measure, Modulation and Metadesign: NC Fabrication in Industrial Design and Architecture," *Journal of Design History*, 11 2019. [Online]. Available: https://doi.org/10.1093/jdh/epz042

[4] G. D. Knott, *Interpolating cubic splines*. Springer Science & Business Media, 2012, vol. 18.

[5] J. O. Ramsay *et al.*, "Monotone regression splines in action," *Statistical Science*, vol. 3, no. 4, pp. 425–441, 1988.

[6] L. Xu, Y. Wang, T. Lux, T. Chang, J. Bernard, B. Li, Y. Hong, K. Cameron, and L. Watson, "Modeling i/o performance variability in high-performance computing systems using mixture distributions," *Journal of Parallel and Distributed Computing*, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0743731519302746

[7] X. He and P. Shi, "Monotone b-spline smoothing," *Journal of the American Statistical Association*, vol. 93, no. 442, pp. 643–650, 1998.

[8] F. Fritsch and R. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis*, vol. 17, no. 2, pp. 238–246, 1980. [Online]. Available: https://doi.org/10.1137/0717021

[9] Q. Wang and J. Tan, "Rational quartic spline involving shape parameters," *Journal of Information and Computational Science*, vol. 1, no. 1, pp. 127–130, 2004.

[10] A. R. M. Piah and K. Unsworth, "Improved sufficient conditions for monotonic piecewise rational quartic interpolation," *Sains Malaysiana*, vol. 40, no. 10, pp. 1173–1178, 2011.

[11] J. Yao and K. E. Nelson, "An unconditionally monotone c2 quartic spline method with nonoscillation derivatives," *Advances in Pure Mathematics*, vol. 8, no. LLNL-JRNL-742107, 2018.

[12] G. Ulrich and L. Watson, "Positivity conditions for quartic polynomials," *SIAM Journal on Scientific Computing*, vol. 15, no. 3, pp. 528–544, 1994. [Online]. Available: https://doi.org/10.1137/0915035

[13] W. Hess and J. W. Schmidt, "Positive quartic, monotone quintic c2-spline interpolation in one and two dimensions," *Journal of Computational and Applied Mathematics*, vol. 55, no. 1, pp. 51–67, 1994.

[14] T. C. H. Lux, L. T. Watson, T. H. Chang, L. Xu, Y. Wang, and Y. Hong, "An algorithm for constructing monotone quintic interpolating splines," in *Proceedings of the High Performance Computing Symposium*, ser. HPC '20. Society for Computer Simulation International, 2020.

[15] T. C. H. Lux, L. T. Watson, T. H. Chang, J. Bernard, B. Li, X. Yu, L. Xu, G. Back, A. R. Butt, K. W. Cameron *et al.*, "Nonparametric distribution models for predicting and managing computational performance variability," in *SoutheastCon 2018*. IEEE, 2018, pp. 1–7.

[16] W. D. Norcott. (2017) Iozone filesystem benchmark. [Online]. Available: http://www.iozone.org

[17] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *Journal of the American Statistical Association*, vol. 62, no. 318, pp. 399–402, 1967.

[18] D.-T. Lee and B. J. Schachter, "Two algorithms for constructing a delaunay triangulation," *International Journal of Computer & Information Sciences*, vol. 9, no. 3, pp. 219–242, 1980.

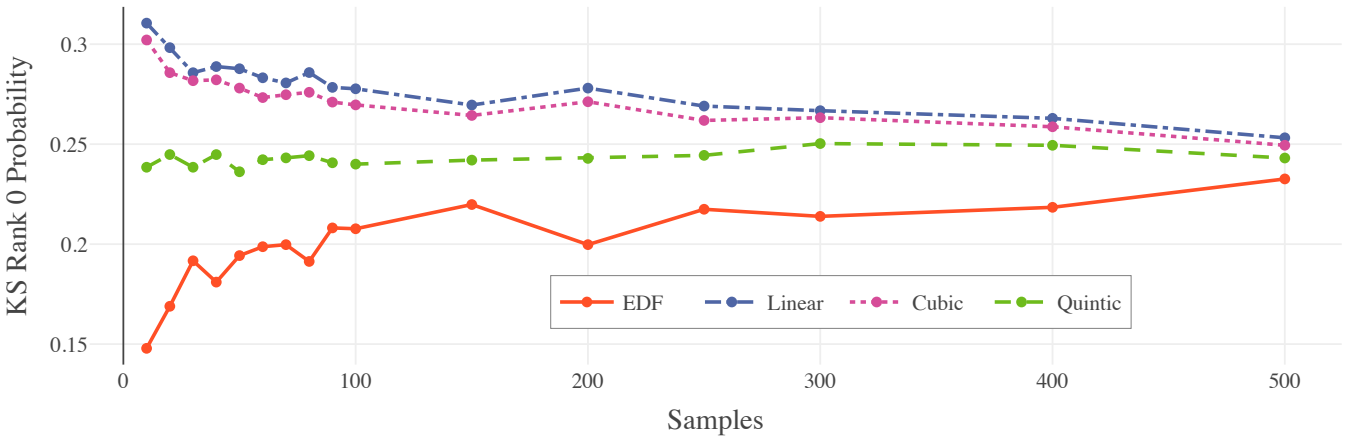

Fig. 6. The probability that any one of the distribution approximation techniques has the lowest KS statistic among all the techniques when given a varying number of samples. Notice that the linear approximation remains the most likely to have the smallest KS statistic for all sample sizes, closely followed by the cubic.