Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

NONPARAMETRIC DENSITY ESTIMATION - A NUMERICAL EXPLORATION

Paul F. Evangelista

Office of Data and Analytics
Taylor Hall
United States Military Academy
West Point, NY 10996, USA

Vikram Mittal

Department of Systems Engineering
Mahan Hall
United States Military Academy
West Point, NY 10996, USA

ABSTRACT

The simulation of distributions without parametric assumptions requires direct estimation of the underlying density function from sample data. Extensive literature discusses the theoretical aspects of this problem. This paper discusses application and practical implications of nonparametric density estimation. Primarily using piece-wise linear interpolation and the Nadaraya–Watson kernel regression methods, tests and experiments show the suitability of nonparametric methods for various circumstances. Nonparametric density estimation has the potential to support complex distributions, which would enable accurate simulation in a fully-automated environment.

1 INTRODUCTION

This paper seeks to answer a fundamental question related to simulation models: When is nonparametric distribution fitting appropriate?

Simulation models derive validity from the accuracy of their representation of real-world systems. When a simulation model accurately represents the essence of a real-world system, typically measured through the comparison of simulation model output statistics and data collected from the real-world system, modelers deem a model as valid (Law 2005). The random variation expected from simulation models typically occurs through random number generation and the transformation of random numbers into various distributions. It is common for modelers to attempt to fit system data to a parametric distribution, often assessed with various goodness of fit tests (chi-square, Kolmogorov-Smirnov, and Anderson-Darling) (Law 2015).

Fitting data to parametric distributions incurs advantages and disadvantages. The advantages include opportunities for analytical or closed-form observations and reasonable extreme event behavior. However, the primary disadvantages involve difficulty in fitting data and limited use with irregular or multi-modal distributions. This paper focuses on the use of empirical distributions to support simulation analysis. While fitting and using parametric distributions to generate simulation models remains an accepted practice, this paper will highlight the cases and thresholds where empirical data serves as a reasonable substitute or improved solution.

2 RELATED WORK

The simulation of random variables garners a long and well-researched body of knowledge. The research presented in this paper numerically and visually explores practical applications related to the simulation of random variates from nonparametric density estimation. Devroye contributed seminal ideas to nonparametric density estimation theory, with particular emphasis in the $L1$ measurement space, which he cites as ”the natural space for densities” due to its invariance and well-defined nature (Devroye and Györfi 1985).

Bratley et al. (1983) provide a summary, theoretical discussions, and algorithms that explore both the first principles of random variate generation as well as advanced topics and unsolved problems, such as tail behavior and small-samples. Izenman (1991) provides a detailed survey of the theoretical underpinnings of nonparametric density estimation and powerful examples of application, both compiling and exploring fundamental ideas such as histograms, kernel methods, and maximum likelihood approaches. Nonparametric density estimation directly contributed to the growth of pattern recognition methods that support machine learning, such as support vector algorithms (Shawe-Taylor and Christianini 2004).

Yücesan (1984) provides a tutorial that presents several algorithms focused on nonparametric methods for distribution estimation. Much of Yucesan's discussion centers on re-sampling techniques, such as bootstrapping the original data and various permutation methods. Yucesan asserts that while the permutation methods in particular are computationally expensive, these techniques yield results that are asymptotically as powerful as test results from parametric tests when parametric assumptions are true. The methods in the present paper offer non-parametric methods that use the structure of the underlying sample to simulate the unknown distribution without relying on re-sampling techniques.

Despite the power and scientific support for nonparametric density estimation, parametric density estimation and distribution fitting finds favor within many contemporary texts that support simulation education (Harrell et al. 2012; Law 2015). Nonparametric density estimation seems to have found its place in automated applications such as machine learning, however the place of nonparametric density estimation within contemporary simulation education is not well-established. This paper offers accesssible ideas and methods, as well as arguments supporting the suitability of nonparametric density estimation for contemporary simulation studies.

3 RANDOM VARIATE GENERATION

3.1 Linear Interpolation

Estimation of a cumulative density function by piece-wise linear interpolation from a sample is a simple exercise. Given an ordered sample, $(X_1, X_2, ...X_n)$, find $F(x)$ by finding X_i and X_{i+1} such that $X_i \leq x \leq X_{i+1}$. $F(x) = i/n + (x - X_i)/(n(X_{i+1} - X_i))$. This interpolation assumes $X_1 \leq x \leq X_n$, which is a reasonable assumption for large n, however for small n, behavior of distribution tails is a well-known challenge which will be discussed later in this paper. Inverse transformation of this function for a uniform random variate, U, yields $x = (U - i/n)(nX_{i+1} - nX_i) + X_i$. Values of $U < i/n$ are either ignored or other methods to estimate tail behavior must be assumed.

3.2 Nadaraya–Watson Algorithm

For small values of n, linear interpolation creates a jagged function which invites various smoothing efforts. The smoothing applied in this paper is the Nadaraya-Watson (NW) kernel regression method (Nadaraya 1964). The NW kernel regression method extends to approximating a cumulative density, $F(x)$, with the following implementation: $\hat{F}(x) = \sum_i iK_h(x - X_i)/n\sum_i K_h(x - X_i)$

The kernel function, K_h, used in this paper is a gaussian kernel. Choice of bandwidth complicates the automation of this algorithm, however there are several known methods for assuming a reasonable estimate for h (Scott and Terrell 1987).

3.3 Exponential Tails

While the aforementioned linear interpolation and kernel regression (or other regression) methods provide reasonable approximation of distributions, approximation of the distribution tails creates challenges. Linear interpolation creates an abrupt truncation of distributions, which may not be appropriate. Weissman (1978) shows that for a broad range of distributions within the exponential class, tail behavior beyond a random point, defined as X_{n-k}, has a tail which can be approximated with an exponential distribution.

Bratley, Fox, and Schrage (1983) provide a lucid explanation and algorithmic implementation of this property. The implementation creates an estimated cumulative density function (CDF), $F(t)$, from linear interpolation or regression for values of $t \leq X_{n-k}$. For values of $t > X_{n-k}$, $F(t) = 1-(k/n)e^{-(t-X_{n-k})/\theta}$, where $\theta = (X_{n-k}/2+\sum_{i=n-k+1}^{n}(X_i - X_{n-k})/k$.

3.4 Suitability of the Nonparametric Density Estimate

Figure 1 provides an example of a population, sample, linearly interpolated density, and NW kernel regression density with exponential tails. The normal distribution provides a useful illustration with a familiar density shape. The linearly interpolated density reflects jagged turns, which are not likely within most underlying distributions, as well as problematic behavior at the tails. Both tails have been estimated with exponential decay as described in section 3.3. Both the interpolated density and kernel density appear to underestimate the distribution, obviously driven by the sample, but the tail behavior appears attractive and reasonable.

Bratley et al. (1983) (p. 123) include a compelling and entertaining discussion of "when not to use a theoretical distribution." They cite the low power of distribution fitting, challenges in parameter estimation, and the computational complexity involved in random number generation from some theoretical distributions. They also provide a power argument related to model sensitivity. If the model is sensitive to the chosen distribution, greater scrutiny is warranted. Providing a contrarian and critical view of using empirical distributions, Harrell et al. (2012) cite several reasons *not* to use an empirical distribution, to include choppiness due to irregularities (linear interpolation) and failing to properly account for extreme values. This criticism merits mentioning as an example of the manner in which some contemporary simulation education commonly dismisses or fails to mention nonparametric density estimation of Harrell et al. (2012) (p. 822). Law (2015) (p. 283–284) provides a much more balanced perspective on the use of empirical distributions, with some theoretical discussion and references, however Law clearly finds favor in simulating theoretical distributions over the use of empirical distributions. The scholarly work of Bratley, Fox, Schrage, and Devroye provide the theoretical underpinnings that support the use of nonparametric density estimation (Bratley et al. 1983; Devroye and Györfi 1985). The experiment that follows attempts to provide further support to reinforce the importance of acknowledging the value of nonparametric density estimation.

4 EXPERIMENTAL METHOD

Consider an experiment that includes a robust population of data that serves as ground truth, a data sample used to estimate a density, and a density estimation strategy. Three methods will be compared to estimate the CDF: linear interpolation of the empirical cumulative density function (ECDF) from the sample; NW kernel regression smoothing of the ECDF from the sample; and direct calculation of the theoretical CDF using the maximum likelihood estimates (MLE) of the parameters from the sample. Exponential tails are included for both the linear interpolation and NW regression methods. The performance measure will be the *p*-value associated with the Anderson-Darling (AD) two-sample statistic, A^2_{nm}, from (Pettitt 1976), which stems from the seminal paper by Anderson and Darling (1954). Pettit's implementation has been implemented in the R package *kSamples* (Scholz and Zhu 2019), and the authors of this paper have implemented and verified the AD two-sample test statistic from the first principles explained in the paper by Pettitt (1976). All experimental analysis and graphics in this paper have been produced in *R* (R Core Team 2022).

The overarching experimental method follows. Given an ordered sample from a known distribution, $Z_1 < Z_2 < ... < Z_l$, estimate the CDF using n equally spaced points between Z_1 and Z_l. l was a factor in the experiment, and n was fixed at 500. From this estimated CDF, generate random variates $X_1 < X_2 < ... < X_n$. $Y_1 < Y_2 < ... < Y_m$ represents the experiment population, created by generating m = 10,000 random variates from the known distribution. For each of the methods examined in this experiment (NW regression, linear

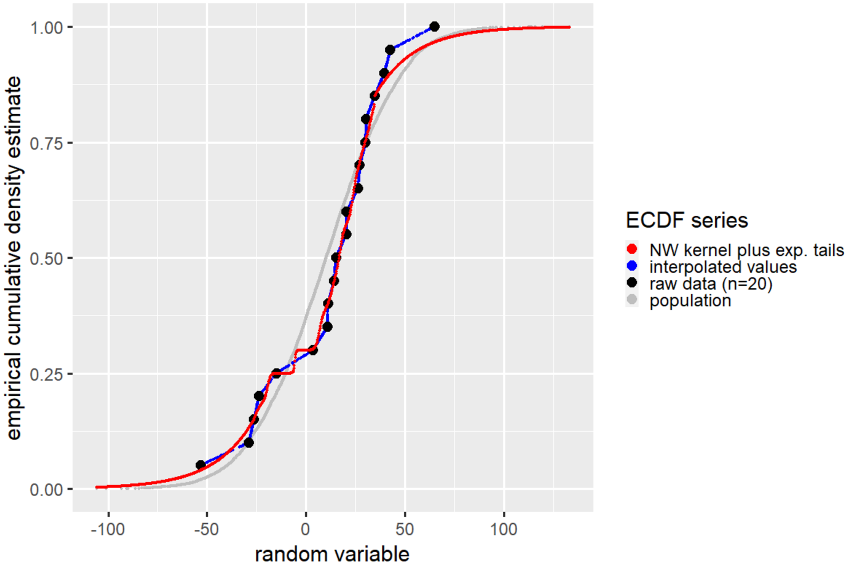


Figure 1: An example of a population, sample, interpolated density, and kernel regression derived density with exponential tails.

interpolation, and the theoretical CDF from MLE), generate $X_1,...,X_n$ and compare to $Y_1,...,Y_m$, which generates A^2_{nm} and an associated p-value. Figure 2 illustrates some aspects of the experimental design.

The factors of this experiment included sample size (l), the bandwidth modifier (s) for the Gaussian kernel used in the NW kernel regression method, and the exponential tail modifier (t). The sample size, l, assumed values within (40, 50, 100, 250, 500, 100, 1000, 2000); s assumed values within (2, 5, 10); and t assumed values within (0.025, 0.05, 0.1). The exponential tail modifier, t, controlled the number of sample points in the tails of the simulated distribution represented by the exponential function. Section 3.3 discusses the algorithm for inclusion of exponential tails. The tail modifier, t, controls the value of k, the exponential tail size, from the algorithm explained in section 3.3. For the experiment, exponential tails were 2.5%, 5%, or 10% of n, the sample size, thus $k \approx tn$. Multiple linear regression calculations revealed that the sample size, l, was the only statistically significant factor. The Gaussian kernel bandwidth was estimated using the unbiased cross-validation function (*bw.ucv()*) from the *stats* package (R Core Team 2022). The calculated bandwidth was divided by s. This reduced the size of the bandwidth, tightening the fit of the NW kernel regression method. Figure 2 clearly shows the result of a tighter fit for the case when $s = 10$.

Notice in Figure 2 how the estimated CDF lines assume a step function behavior when $s = 10$. For small bandwidths, the weighting of the kernel function focuses almost exclusively on the closest points, creating this step behavior. As the bandwidth increases, smoothing increases as the weight in the function distributes to other points nearby. Figure 2 also clearly shows the effect of increased l. It is intuitive that larger samples improve the estimates; experimentation validates this intuition and provides insight into rates of improvement as l grows.

5 RESULTS

The measure of success for the experiment conducted in this paper is defined as $P(A^2_{nm} > \hat{A}^2_{nm})$, where A^2_{nm} is the AD two-sample random variable and $\hat{A}^2_{nm}$ is the measured AD two-sample statistic. Let $\bar{p}_k$, $\bar{p}_l$, $\bar{p}_t$ represent the average p-value of the AD two-sample statistic based on 30 iterations for each design point, using the NW kernel method, linear interpolation, and MLE theoretical distribution, respectively. Low p-values for distribution-fitting tests indicate a poor fit; higher values indicate that there is no reason

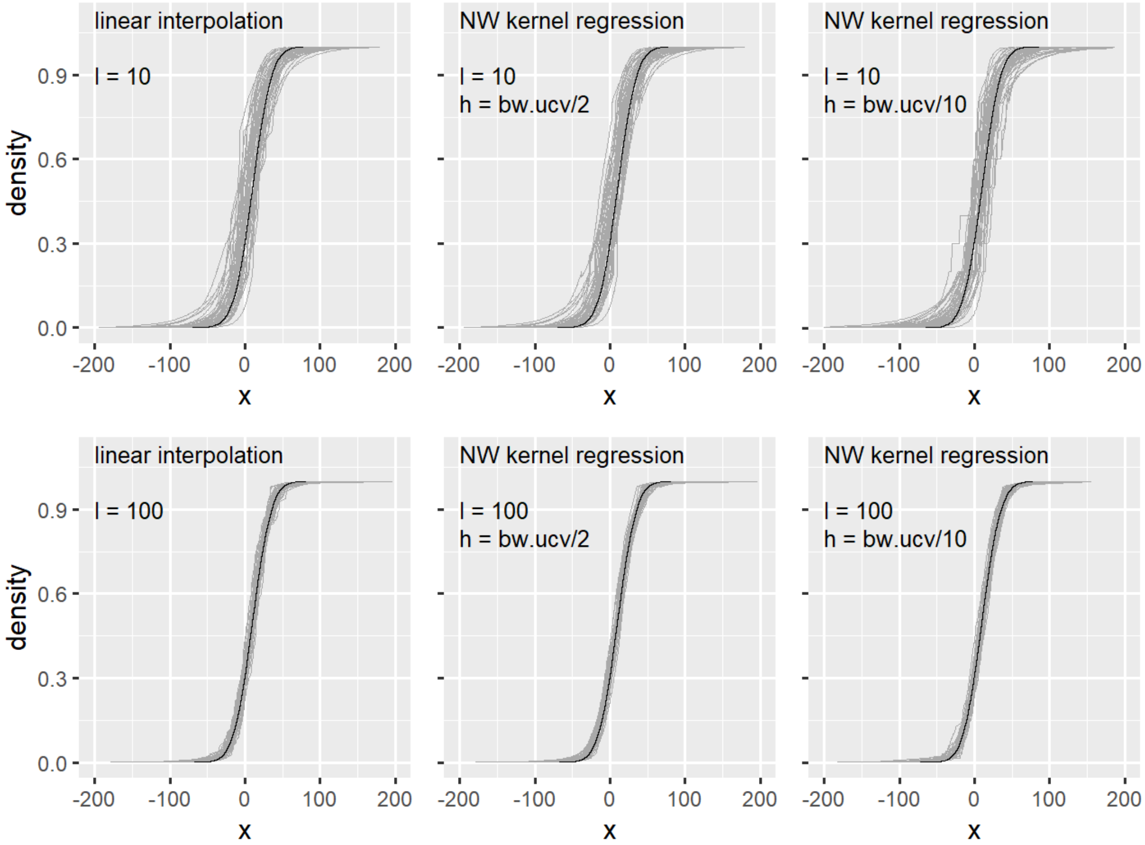


Figure 2: Visualizing NW kernel and linear interpolation density estimation for some of the design points.

to reject the null hypothesis, where the null hypothesis assumes that the distribution of the simulated data equals the distribution of the underlying population. Additionally, $\bar{p}_t - \bar{p}_k$ was also measured, in order to directly assess the relationship between the MLE theoretical distribution and NW kernel regression estimated distribution. The known distribution assessed in Table 1 was the normal distribution.

Table 1 shows that the sample size, l, was the only statistically significant factor in the experiment. The estimated effect of l is a positive fraction, with error showing that it is clearly greater than zero. This positive fraction indicates that as the sample size grows, the p-value grows. This means that as the sample size grows, the likelihood of estimating an acceptable distribution from the non-parametric methods described in this paper also grows. The size of the exponential tails and the bandwidth modifier for the Gaussian kernel did not have a statistically significant effect.

Several insights emerge from these results. As expected, estimation of a CDF with MLE parameters fits the theoretical distribution well. It should. Theoretical estimates of distributions with MLE parameters outperform nonparametric estimation for small sample sizes. As the sample size grows, the fit of nonparametric density estimation improves and ultimately performs comparably with parametric density estimation. It is important to realize that when comparing nonparametric density estimation with parametric density estimation within this experiment, this has been done *with knowledge* of the underlying distribution, or a huge population sample to accurately estimate this underlying distribution. In practice, this is often not the case. As Bratley et al. (1983) (p. 124) eloquently state, “God does not usually tell us from what distribution the data come.”

Table 1: Regression table showing the effect of sample size (l), presence of exponential tail algorithm (t), and bandwidth modifier (s).

	Dependent variable:		
	$\bar{p}_k$		$\bar{p}_t - \bar{p}_k$
	(1)	(2)	(3)
l	0.0002***	0.0002***	−0.00005***
	(0.00001)	(0.00001)	(0.00001)
t	−0.035		−0.208
	(0.298)		(0.208)
s	−0.0005		−0.002
	(0.003)		(0.002)
Constant	0.066**	0.061***	0.136***
	(0.027)	(0.012)	(0.019)
Observations	63	63	63
R^2	0.800	0.800	0.298
Adjusted R^2	0.790	0.796	0.262
Residual Std. Error	0.074 (df = 59)	0.073 (df = 61)	0.051 (df = 59)
F Statistic	78.549*** (df = 3; 59)	243.423*** (df = 1; 61)	8.350*** (df = 3; 59)
Note:			*p<0.1; **p<0.05; ***p<0.01

Furthermore, when considering a seemingly endless array of future automation, human intervention with distribution fitting is likely to become impractical. Additionally, odd-shaped distributions and multi-modal behaviors may not fit any parametric distribution. Nonparametric density estimation has a place within simulation modeling, and it arguably has a much more prominent place in contemporary simulation modeling education and curriculum.

Figure 4 shows several additional comparisons between the fit of kernel regression estimates and linear interpolation estimates. Figure 4(i) shows the effect of removing the exponential tail estimates. The NW kernel regression estimated distributions clearly begin to outpace the linear interpolation method, presumably due to the effect of sporadic tail behavior on a piece-wise linear interpolation. If it is possible to remove the tails altogether and simply use the NW regression method, which appears to be a viable method, this eliminates unnecessary complexity and modeling parameters. Figure 4(ii) shows the results of estimating a bimodal distribution. Assume Z_1 and Z_2 represent two normally distributed random variables, independent and distributed with a mean of μ_1 and μ_2, respectively, and equal variance of σ^2. B represents a Bernoulli random variable with $p = 0.5$. W will be bimodal when $W = BZ_1 + (1-B)Z_2$, assuming the difference between the means of Z_1 and Z_2 are separated by $\tilde{2}\sigma$ or more. Figure 4(ii) shows the results of estimating the distribution of W. Figure 4(iii) uses a similar method to create a multi-modal distribution, introducing a second bernoulli variable. A multi-modal distribution with four modes results from $W = B_1Z_1 + B_2Z_2 + Z3$. If $(\mu_1, \mu_2, \mu_3) = (100, 200, 1)$, with $\sigma = 1$, a distribution with four modes emerges. Figure 4(iii) shows the results of fitting this distribution. Lastly, Figure 4(iv) shows the results of fitting a distribution that contains the monthly fractional change of the S&P500 stock market index from the year 1871 to 2018 (Shiller 2022). This market data shows some symmetry with a bell-shaped density, however it does not fit a normal distribution well.

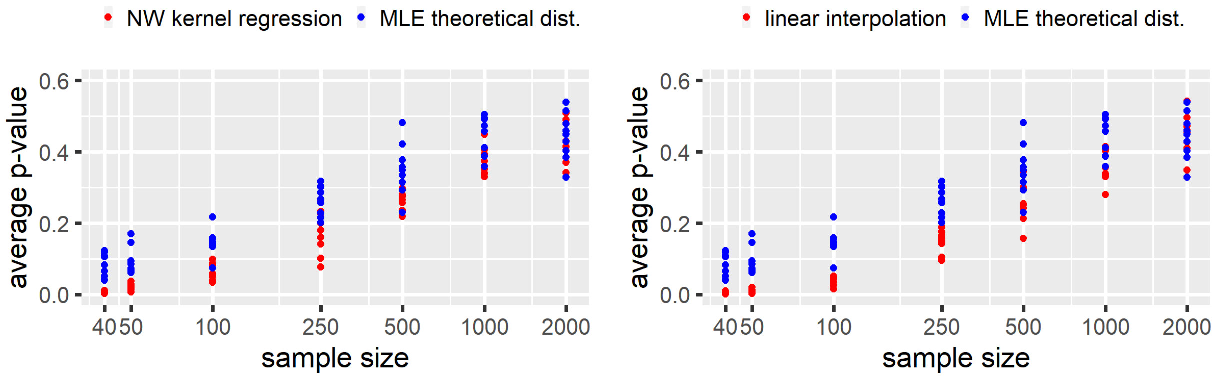
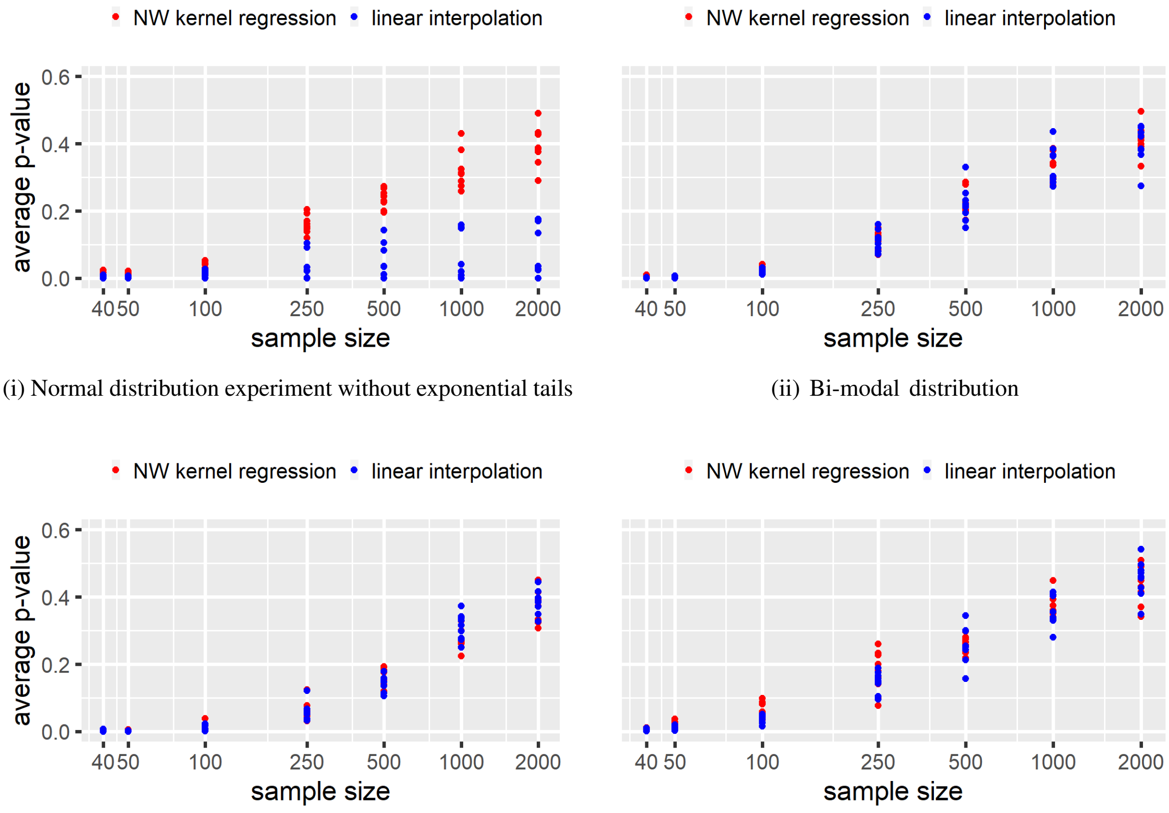


Figure 3: The average p-value of A^2_{mn} was used to measure the fit of an estimated distribution compared to a normal distribution. Distribution estimate methods included NW kernel regression with exponential tails, linear interpolation with exponential tails, and the theoretical fit using MLE parameters.



(i) Normal distribution experiment without exponential tails

(ii) Bi-modal distribution

(iii) Multi-modal distribution

(iv) SP500 data

Figure 4: Experimentation with additional distributions and data.

6 FUTURE WORK AND CONCLUSION

The use of empirical distributions can be useful for a number of simulation applications. One area would be combat simulations, which rely heavily on Monte-Carlo simulations for determining the accuracy of shooting data, whether they be rifles or tanks. Most combat simulations rely on normal distributions to capture the variation in shot location; however, in actuality, the data would not follow a standard distribution (Tolk 2012). Ample shooting data is collected annually by militaries as part of marksmanship and gunner training. Using this data to drive an empirical distribution would provide for more robust and accurate simulations.

Another potential applications is related to driving behavior in electric and hybrid vehicles. Future vehicles will include control systems that reflect the behavior of the driver using predictive analysis. This data will certainly not fit a common distribution due to the large degree of variability of possible drivers. The use of empirical in these applications could support the models that underly the control systems (Ling et al. 2020). A similar strategy would also be applicable for the underlying simulations that support the predictive analysis required for optimizing power grids (Quan et al. 2014).

Nonparametric density estimation deserves a role in simulation, particularly in simulation education. As the amount of available data and automated methods of collecting data grows, the viability of nonparametric density estimation also grows. The community of educators and simulation have not rendered a clear judgement on the use of nonparametric density estimation, showing that this topic deserves additional experimentation and theoretical research.

REFERENCES

Anderson, T. W., and D. A. Darling. 1954. “A Test of Goodness of Fit”. *Journal of the American Statistical Association* 49(268):765–769.

Bratley, P., B. L. Fox, and L. E. Schrage. 1983. *A Guide to Simulation*. New York: Springer Verlag.

Devroye, L., and L. Györfi. 1985. *Nonparametric Density Estimation : The L[1] View*. New York: John Wiley Sons.

Harrell, C., B. Ghosh, and R. Bowden. 2012. *Simulation Using ProModel*. McGraw-Hill Series in Industrial Engineering and Management Science. McGraw-Hill/Higher Education.

Izenman, A. J. 1991. “Recent Developments in Nonparametric Density Estimation”. *Journal of the American Statistical Association* 86(413):205–224.

Law, A. M. 2005. “How to Build Valid and Credible Simulation Models”. In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. Kuhl, N. Steiger, F. Armstrong, and J. Joines, 24–32. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Law, A. M. 2015. *Simulation Modeling and Analysis*. 5th ed. New York: McGraw Hill.

Ling, K., N. Shah, and J. Thiele. 2020. “Customer-Centric Vehicle Usage Profiling Considering Driving, Parking, and Charging Behavior”. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. September 20th-23rd, Rhodes, Greece, 1-6.

Nadaraya, E. A. 1964. “On Estimating Regression”. *Theory of Probability & Its Applications* 9(1):141–142.

Pettitt, A. N. 1976. “A Two-Sample Anderson–Darling Rank Statistic”. *Biometrika* 63(1):161–168.

Quan, H., A. Khosravi, D. Yang, and D. Srinivasan. 2014. “Incorporating Wind Power Forecast Uncertainties into Stochastic Unit Commitment Using Neural Network-based Prediction Intervals”. *IEEE Transactions on Neural Networks and Learning Systems* 26(9):2123–2135.

R Core Team 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Scholz, F., and A. Zhu. 2019. *kSamples: K-Sample Rank Tests and their Combinations*. R package version 1.2-9.

Scott, D. W., and G. R. Terrell. 1987. “Biased and Unbiased Cross-Validation in Density Estimation”. *Journal of the American Statistical Association* 82(400):1131–1146.

Shawe-Taylor, J., and N. Christianini. 2004. *Kernel Methods for Pattern Analysis*. New York: Cambridge University Press.

Shiller, R. 2022. “S-and-P-500”. https://github.com/datasets/s-and-p-500, accessed June 15th.

Tolk, A. 2012. *Engineering Principles of Combat Modeling and Distributed Simulation*. New York: John Wiley and Sons.

Weissman, I. 1978. “Estimation of Parameters and Larger Quantiles Based on the k Largest Observations”. *Journal of the American Statistical Association* 73(364):812–815.

Yücesan, E. 1984. “Nonparametric Techniques in Simulation Analysis: A Tutorial”. In *Proceedings of the 1994 Winter Simulation Conference*. December 11th-13th, Orlando, Florida.

AUTHOR BIOGRAPHIES

PAUL EVANGELISTA is a Colonel in the United States Army and an associate and academy professor at the United States Military Academy (USMA). He is currently serving as the USMA chief data officer and teaches in the Department of Systems Engineering. His email is paul.evangelista@westpoint.edu. His website is http://paul-evangelista.com.

VIKRAM MITTAL is an Associate Professor in the Department of Systems Engineering at the United States Military Academy. He has taught courses in both the systems engineering and mechanical engineering departments. Prior to teaching at USMA, he was a senior mechanical engineer at the C.S. Draper Laboratory, where he worked in the Vehicles and Robotics Group. Dr. Mittal’s research interests focus on combat simulations, model-based systems engineering, robotics, and power systems. His e-mail is vikram.mittal@westpoint.edu.