Generalizing Distribution Estimation To Form Fitting

**M.C. Rumuly**  
ECEN 662-600  
Texas A&M UniversityCollege Station, TX

*Abstract*—This is the sound of speed

Keywords—do, I, even, need, to, care, about, this, section

# Introduction

Characterizing unknown distributions is important for analyzing and modeling random systems. Given a sample, focused on the problem of estimating the parameters of a known distribution. Want to generalize to the case where the underlying form of the distribution is unknown. For this case want to find the distribution of the common ones which most accurately models the underlying system. Move beyond histograms and human intuition

Focus on continuous random variables here. Want a ‘goodness of fit’ metric which can be used to compare different distributions. Look at as a frequentist problem for now; Bayesian analysis is beyond the scope of this project. Only seeks to characterize the sample as belonging to one of a number of known, defined distributions with estimated parameters rather than attain arbitrary accuracy; maximum accuracy is always at the empirical distribution itself, which cannot be used for analysis of the sort a more general characterization is useful for.

General comparison of distributions; describe EDF and relation to CDFs (and consequential relation between sample and PDF).

# Objective Functions

Two models dominate the methods of comparing distributions: the first is based in Shannon Entropy, from information science. The second is based on comparing the cumulative distribution functions. For an empirical sample, the cumulative distribution function is called the Empirical Distribution Function (EDF). (NOTE: include criterion formula)

## KullBack-Leibler Divergence

Uses Entropy. Would be remiss not to mention, but doesn’t work for comparing sample (EDF not smooth) and continuous distribution. As such, excluded from testing in this domain. [CITATION NEEDED]

## Kolmogorov-Smirnov Test

Uses Maximum Error between EDF and CDF. Proven convergence to 0. [CITATION NEEDED]

## Kuiper’s Test

Refinement on Kolmogorov-Smirnov, equally sensitive to tails and median [CITATION NEEDED]

## Anderson-Darling Test

Based on mean square error, basic statistic places most of weight on tails, logarithm makes undefined at tails (where CDF becomes 0 or 1). Hard to use. Capable of rejecting distribution based on critical value dictionary. [CITATION NEEDED]

## Cramer-von Mises Criterion

Refinement on Anderson-Darling. Does not require logarithm; more useful/calculable. Also capable of rejecting based on critical value dictionary. [CITATION NEEDED]

# 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

## Uniform Continuous

Most basic, ubiquitous, often used for computing random numbers in video-games etc. Equal probability weight over finite interval.

## Normal

Perhaps most useful in statistical analysis due to additive properties. Bell-curve.

## Exponential

Often used to model arrival times. Single-tailed, time invariance.

## Laplace

Useful for difference between exponential arrival times, but most useful here because visually similar to normal distribution, while still clearly distinct.

Normal, Uniform, and Exponential chosen for ubiquity, Laplace chosen specifically to muddle (especially with Normal).

# Test Method

For a given real distribution Fr(x), generate random sample X, X = (x1, x2, … xN). Now, for each distribution Fe(x), estimate parameters (invalidating uneeded rejection capability of Anderson-Darling and Cramer-von Mises), calculate each statistic, compare. Minimum is winner for that. Record correctness. Average statistic and correctness across trials. For each sample size N in {list}. Stick equations here.

Implementation scalable, included in appendix [APPENDIX]

# Results

Accuracy of each statistic in identifying each distribution

Behavior of each objective function with each set of distributions

# Conclusion

Success of this implementation proves concept; note that this implementation can be easily adapted to take generic input and return closest model.

Statistic by statistic: general usefulness

Future work: Bayesian extension, rigorous confidence proof beyond empirical study.

##### References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.