Performance of Maximum Entropy Probability Density in the Case of Data Which are not Well Distributed

ALADDIN SHAMILOV*
ILHAN USTA*
YELIZ MERT KANTAR*
*Anadolu University
Science and Art Faculty-Department of Statistics
16470-Eskisehir
TURKEY

Abstract: - Maximum entropy (MaxEnt) principle is a method for analyzing the available information in order to determine a unique probability distribution. Generally, this information is given as moment constraints of system. Obtained distribution is called MaxEnt probability density function. In this study, it is shown the excellent agreement of the MaxEnt densities with data which are not modelled by the known statistical distributions. On this purpose, special data are simulated. The distribution of this data is obtained by MaxEnt method. The agreement of data to MaxEnt density function is shown various statistical criteria.

Key words:- Maximum Entropy principle, MaxEnt probability density function, Correlation coefficient, Chi-square, Root mean square error.

1 Introduction

Recently the methods based on information theory to obtain the distribution of random variable are very popular. The maximum entropy (MaxEnt) [1] and the minimum cross entropy (MinxEnt) are such of these methods [2], [3], [4]. There have been a lot of studies about the application of these methods in literature. In [5], MaxEnt method was employed to approximate the size distribution of U.S. family income and MaxEnt distributions were compared with two conventional income distributions. In [6], MaxEnt models are used to determine the distribution of diurnal, monthly, seasonal and yearly wind speed.

The rest of the paper is organized as follows. Section 2 introduces MaxEnt method. In section 3, in order to show the excellent agreement of the MaxEnt densities with data which are not modelled by the known statistical distributions, performance studies are done. In here, data are simulated so as to not fit statistical distributions. Also various criteria are given to evaluate the performance of MaxEnt distributions. Section 4 concludes the paper with some suggestions for of further research

2 MaxEnt Method

The Principle of Maximum Entropy is a method that can be used to estimate input probabilities more generally. The result is a probability distribution that is consistent with known constraints expressed in terms of averages, or expected values. MaxEnt is the most unbiased or most uniform probability distribution conditioned upon the available information [7]. The MaxEnt approach covers a whole family of generalized exponential distributions, including the exponential, Pareto, normal, lognormal, gamma, beta distribution as special cases. For example, if the mean (first moment) is prescribed, then the MaxEnt density is exponential.

The MaxEnt density function of continuous random variable X is obtained by maximizing Shannon's entropy measure

$$H = -\int_a^b f(x) \ln f(x) dx \tag{1}$$

subject to the constraints:

$$\int_a^b f(x) dx = 1 \tag{2}$$

$$E\{g\} = \int_a^b g_k(x) f(x) dx = \mu_k \quad (3)$$

where $g_k \ (k = 1, \ldots m)$ is moment vector function, $\mu_k \ (k = 1, \ldots m)$ is moment vector value.

Obtained MaxEnt density function is

$$f(x) = \exp(-\lambda_0 - \sum_{k=1}^{m} \lambda_k g_k(x)) \ .$$

where $\lambda_0, \lambda_1, \ldots, \lambda_k$ are the Lagrange multiplier for the ith moment constraint. Since an analytical solution does not exist for $k \geq 2$, one must use a nonlinear optimization technique to solve for the MaxEnt density. These Lagrange multipliers can be found by Newton method. In literature, there have been numerous studies to calculate these multipliers as [21-25]. But we use the program on MATLAB and use MATLAB's optimization toolbox to compute Lagrange multipliers.

3 Performance Studies

In this section, we test the performance of MaxEnt function in fitting data. On this purpose, we simulate data which are not well distributed and find MaxEnt density of this data. In order to show performance of MaxEnt probability distributions, we use various statistical criteria. Root mean square error (*RMSE*) [8], Chi-square (X^2), Correlation coefficient (R^2) will be used in statistically evaluating the performance of MaxEnt distributions.

The formula of mentioned suitability judgment criteria are

$$RMSE = \left(\frac{\sum_{i=1}^{N} (y_i - x_i)^2}{N} \right)^{\frac{1}{2}}, \quad (15)$$

$$\chi^2 = \frac{\sum_{i=1}^{N} (y_i - x_i)^2}{N - n}, \quad (16)$$

$$R^2 = \left(1 - \frac{\sum_{i=1}^{N} (y_i - x_i)^2}{\sum_{i=1}^{N} (y_i - z_i)^2} \right), \quad (17)$$

where y_i is the ith probability of actual data, x_i is the ith predicted probability, N is number of all observed wind speed data, n is the number of parameters or the number of constrains.

The best distribution function can be determined according to the lowest values *RMSE*, X^2 the highest values R^2.

Histograms of simulated data are given in Figure 1, Figure 2, and Figure 3. From these Histograms, it can be seen that data are not well distribution.

Figure 1 Histogram of simulated data

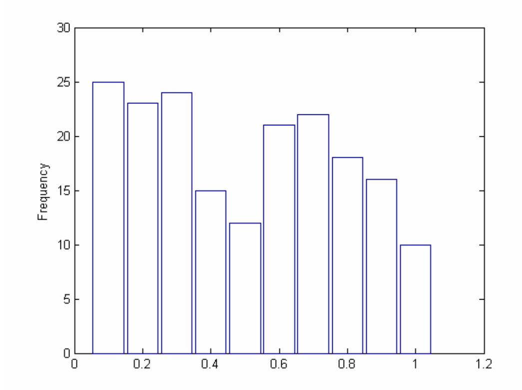


Figure 2 Histogram of simulated data

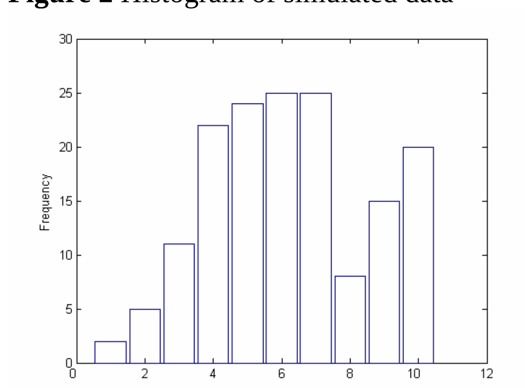


Figure 3 Histogram of simulated data

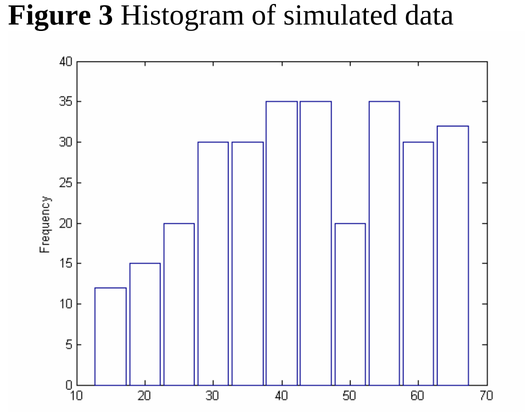


Table 1 The agreement of first data to MaxEnt probability density

	Statistical Analysis	
	$f_{observed}$	f_{MaxEnt}
	0.1344	0.1341
	0.1237	0.1280
	0.1290	0.1144
	0.0806	0.0908
	0.0645	0.0838
	0.1129	0.0918
	0.1183	0.1065
	0.0968	0.1114
	0.0860	0.0900
	0.0538	0.0493
RMSE	0.01242	
X^2	0.00038562	
R^2	0.78173	

Table 2 The agreement of second data to MaxEnt probability density

	Statistical Analysis	
	$f_{observed}$	f_{MaxEnt}
	0.0127	0.0127
	0.0318	0.0331
	0.0701	0.0679
	0.1401	0.1317
	0.1529	0.1753
	0.1592	0.1595
	0.1592	0.1180
	0.0510	0.0893
	0.0955	0.0861
	0.1274	0.1265
RMSE	0.019601	
X^2	0.00096052	
R^2	0.86178	

Table 3 The agreement of third data to MaxEnt probability density

	Statistical Analysis	
	$f_{observed}$	f_{MaxEnt}
	0.0408	0.0413
	0.0510	0.0481
	0.0680	0.0733
	0.1020	0.0975
	0.1020	0.1098
	0.1191	0.1106
	0.1191	0.1059
	0.0680	0.1011
	0.1191	0.0994
	0.1020	0.1022
	0.1088	0.1109
RMSE	0.012967	
X^2	0.000369	
R^2	0.77491	

Normally, a value higher than 70% of R^2 is acceptable [8]. The parameters for the statistical analysis: the R^2, RMSE and Chi-square error are given in Table1, Table 2 and Table 3 for all data. It can be seen that the R^2 vary from 0.77% to 0.86% for all data if MaxEnt density functions subject to five moment constraints (f_{MaxEnt}**)** are used. Also from these tables, f_{MaxEnt} shows better fitting in terms of RMSE, X^2. The other words, Table1-3 and Figure 4-6 shows the excellent agreement of the MaxEnt densities with wind data. For this reason the MaxEnt distributions are much convenient for data is not well distributed.

Figure 4, 5, 6 reports the histogram of the simulated data and the estimated MaxEnt densities with one, two, three, four, five constraints. From these Figures, the obtained MaxEnt densities demonstrates good fitting. Moreover, the MaxEnt distributions are getting better while the number of constraint is added.

Figure 4 Histogram and Best-fit-MaxEnt density curves for firth simulated data

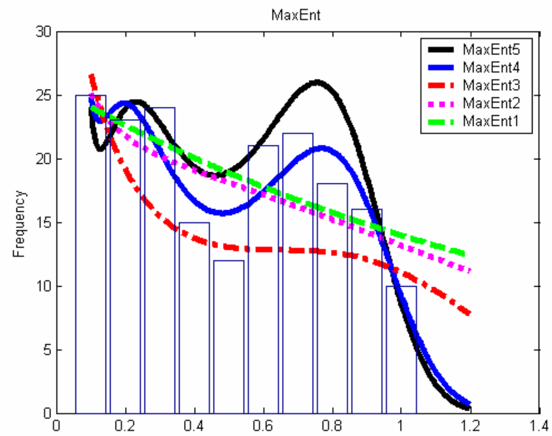


Figure 5 Histogram and Best-fit-MaxEnt density curves for second simulated data

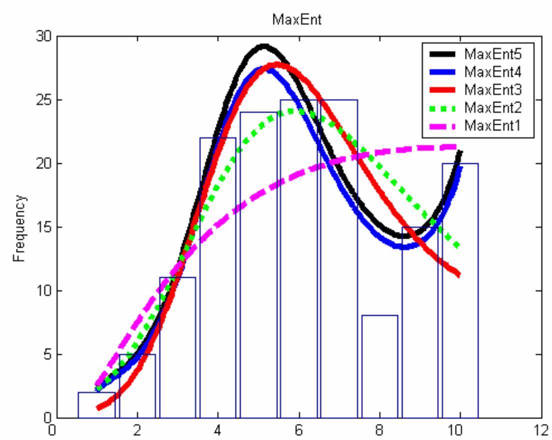
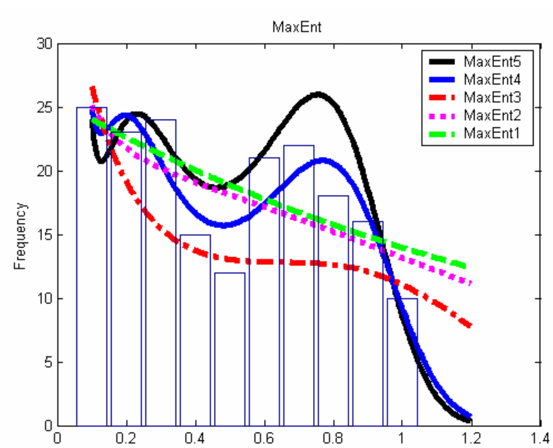


Figure 5 Histogram and Best-fit-MaxEnt density curves for third simulated data



4 Conclusions

The maximum entropy method is a flexible and powerful tool for density approximation. Empirical evidence demonstrates the efficiency of this method. For this reason the MaxEnt distributions are much convenient for data is not well distributed.

References:

[1] Jaynes, E. T. Papers on Probability, Statistics, and Statistical Physics. Netherlands: D. Reidel, Dordrecht, 1983.

[2] S. Kullback and R. A. Leibler, On Information and Sufficiency, Ann. Math. Stat., Vol.22, pp. 79-86, 1961

[3] Kullback, S. *Information Theory and Statistics*. Wiley, 1959.

[4] Kapur, J.N. and Kesavan, H. K. (1992), *Entropy Optimization Principles with Applications.* London: Academic Press, Inc.

[5] Wu, X. Calculation of maximum entropy densities with application to income distribution, *Journal of Econometrics*, 2003, 115, 347-354.

[6] Li M., Li X. Investigation of wind characteristics and assessment of wind energy potential for Waterloo region, *Energy conversion management*, 46, 2005, 3014-3033

[7] Srikanth M, Kesavan H, Peter H. Probability density function estimation using the minmax measure. IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, 2000, Vol.30, No.1, pp.77-83

[8] Celik A.N. A statistical analysis of wind power density based on the Weibull and Rayleigh models at the southern region of Turkey, *Renewable energy*, 29, 2003, 593-604.