American Journal of Mathematics and Statistics 2021, 11(3): 61-66
DOI: 10.5923/j.ajms.20211103.02

Maximum Entropy Estimation: Agronomic Dataset

Raymond Bangura[1,*], Sydney Johnson[2]

[1]Biometric Unit, Sierra Leone Agricultural Research Institute, Freetown, Sierra Leone
[2]Department Agronomy, Rokupr Agricultural Research Center, Kambia District, Sierra Leone

Abstract We provide a generalized maximum entropy method and its application to the Agronomic dataset. Errors of deviation are shown with the analysis of variance on the entropy method. Multi-collinearity is one of the major problems of regression analysis. Based on the maximum entropy method, we gave a better estimate than the traditional robust regression of four independent variables from an Agronomic dataset. From the generalized maximum entropy method, we showed the relationship between the dependent and independent variables. Also, we provided a diagnostic fit for the dependent variable to support the theoretical analysis.

Keywords General maximum entropy, Multicollinearity, Support space, Regression, Residual

1. Introduction

The maximum entropy method has gained a spectrum of recognition in many disciplines, such as Mathematics, Engineering sciences, and statistics. The maximum entropy principle was proposed by Shannon, 1948, which states that an inference is made based on incomplete information draws from the probability distributions that maximizes the entropy that is subject to constraints on the distribution. In order words, when given a large number of probability distributions, we choose the one that best represents the present state. The process of choosing the highest uncertainty is known as the maximum entropy. There are infinite number of possible models that satisfy the constraints (Simon Haykin, pg481). As a result of this, the maximum entropy is a constrained optimization technique. The maximum entropy includes numerical solutions of stationary density functions by Perron-Frobenius operator, introduced by J. Ding 1995 in the study of dynamic systems. A generalized maximum entropy estimator is a robust estimator that resists multicollinearity problems. The maximum entropy estimates the parameters in a linear regression model, especially when the data are ill-posed.

The generalized maximum entropy estimator has played a vital role in the econometric model estimation because it is an alternative estimator to least squares (Wilawan Srichaikul, 2018). Akdeniz et al., 2011, provided an alternative estimation method of maximum entropy to estimate parameters in a linear regression model especially when the basic data are ill-conditioned. The generalized maximum entropy helps find information about variables or measures through probability functions using the Shannon method of general maximum entropy. Based on the generalized maximum entropy method, Bangura et al., 2020, provided a good estimate of three variables from a rice seed data.

To this present moment, the estimation of parameters by the maximum entropy method is still rare because regression analyses are posing with threat of multi-collinearity and ill-conditioned dataset. The primary task in using the GME approach as an estimator to choose an appropriate entropy measure that reflects the uncertainty (state of knowledge) that we have about the occurrence of a collection of events (Wilawan Srichaikul, 2018). The paper also estimates the value of the generalized maximum entropy residuals of the weighted variables, which quantifies the amount of information in a variable, providing the basis for an assumption about the concept of information derived from the variables (weighted). We also found a linear regression from the generalized maximum entropy procedure to show the relationship between endogenous (explained variable) and exogenous (regressors) because a generalized maximum entropy estimator is a robust estimator that is strictly resistant to multicollinearity.

In this paper, we considered four parametric variables are; panicle (Pan), plant height (Plant_Ht), panicle length (Pan_Length), and Tillers (dependent) taken from the data set in 2014 and 2015. The paper is divided into five sections. We introduced the topic in section one. In section we gave the theoretical background of the maximum entropy method followed by its application in section three. We examined the residuals in section four and gave our conclusion in section five.

* Corresponding author:
mannahray@yahoo.com (Raymond Bangura)
Received: Apr. 4, 2021; Accepted: May 7, 2021; Published: May 26, 2021
Published online at http://journal.sapub.org/ajms

2. Generalized Maximum Entropy

Jaynes (1957) proposed the method of maximizing entropy by recovering unknown probabilities by characterize dataset, subject to the available sample moment information, and adding up constraints on the probabilities (Simon Haykin, pg481). $H(p) = -\sum_{i=1}^{n} p_i ln p_i$

Linear Constraint

$$\sum_{i=1}^{n} p_i g_r = a_r$$

For $r = 1, \dots, m$

Within the classic ME framework, the observed moments are assumed to be exact. To extend this approach to the problem with noise, the GME approach (developed by Golan, Judge, and Miller (1996)) generalizes the ME approach by using a dual objective (precision and prediction) function.

The generalized maximum is the covariance estimate shows 596 convergence criteria. It gives the estimate of four parameters. From the classical general linear model (GLM), we have;

$$y_t = \sum_{i=1}^{k} X_{tk} + \beta_k + \mu_{t\,,} t = 1, 2, \dots, N \qquad (1)$$

We choose to define the dimensional vector M with equal distance discrete points. The randomness is called the support space $Z'_1 = [Z_{K1}, Z_{K2}, \dots, Z_{KM}]$ and the vector of probabilities associated with the M dimension is given as $[p_{K1}, p_{K2}, \dots, p_{KM}]$ Now, we can rewrite matrix form as;

$$y = X\beta + \mu \qquad (2)$$

We re-parameterized the unknowns which are; β and μ from (2) in such a way that they both represent probabilities accordingly. That is;

$$\beta = Zp$$

$$\mu = Vw$$

So we have; under this re-parameterization, the inverse problem with noise given in (1) may be rewritten as

$$y = XZp + Vw \qquad (3)$$

X is the $T \times K$ known matrix of explanatory variables and μ is a $T \times l$ noise (disturbance) vector.

According to Golan, 1996 we aim to convert each parameter β_k. If $M \geq 2$ with an equal distance discrete support values, z_{km}, with corresponding probabilities p_{km},. By this way, each parameter is converted from the real line into a well-behaved set of proper probabilities defined over the supports.

$$\beta = zp = \begin{bmatrix} z'_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & z'_K \end{bmatrix} \begin{bmatrix} p_1 \\ \vdots \\ p_K \end{bmatrix}$$

$$\beta_k = z'_k p_k = \sum_{m-1}^{M} z_{km} p_{km}, for\ k = 1, 2, \dots, K$$

$$and\ m = 1, 2, \dots., M$$

Also, the disturbance (noise) μ with assign probabilities $v'_t = [v_{t1}, v_{t2}, \dots, v_{tJ}]$; given that $J \geq 2$. We have;

$$\mu = Wv = \begin{bmatrix} v'_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_T \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_T \end{bmatrix}$$

Likewise;

$$\mu_t = v'_t w_t = \sum_{j-1}^{J} z_{tJ} p_{tJ}, for\ t = 1, 2, \dots, T$$

$$and\ j = 1, 2, \dots., J$$

As for the determination of support bounds for disturbances, Golan et al (1996) recommend using the "three-sigma rule" of Pukelsheim (1994) to establish bounds on the error components: the lower bound is $V_L = -3\sigma_y$ and the upper bound is $V_U = 3\sigma_y$, where σy is the (empirical) standard deviation of the sample y. For example if J= 5, then $'v'_t = (-3\sigma_y, -1.5\sigma_y, 0, 1.5\sigma_y, 3\sigma_y)$ can be used.

Jaynes (1957) demonstrates that entropy is additive for independent sources of uncertainty. Therefore, assuming the unknown weights on the parameter and the noise supports for the linear regression model are independent, we can jointly recover the unknown parameters and disturbances (noises or errors) by solving the constrained optimization problem of max;

$$H(p, w) = -plnp - wlnw$$

Subject to

$$y = XZp + Vw$$

When β_k and μ_k are re-parameterized, and are transformed with assigned probabilities.

$$\max_{p,w} H(p, w) = -\sum_{k=1}^{K} \sum_{m=1}^{M} p_{km}\, ln p_{km} - \sum_{t=1}^{T} \sum_{j=1}^{J} w_{tj}\, ln w_{tj}$$

Constraints

$$\sum_{k=1}^{K} \sum_{m=1}^{M} x_{tk} z_{km} p_{km} + \sum_{j=1}^{J} w_{tj} v_{tj} = y_t\ for\ t = 1, 2, \dots., T$$

$$y_t = \sum_{k=1}^{K} x_k \sum_{m=1}^{M} z_{km} p_{km} + \sum_{j=1}^{J} w_{tj} v_{tj}\ \ for\ t = 1, 2, \dots, T$$

By solving the first order condition, the estimates of the GME parameters are given by;

$$\hat{\beta}_{GME} = Z\hat{p}$$

$$\hat{\varepsilon}_{GME} = V\hat{w}$$

3. Generalized Maximum Entropy Application

3.1. Final Information Measures

Table 1 shows the final information summary as assumed that the information is incomplete in estimating the generalized maximum entropy. The objective function value includes prediction and precision given at 5.49. The objective value function represents the value of the entropy estimation problem. The signal entropy is 6.14, noise is -0.65, normed (signal) is 0.95, normed (noise) is 0.99, parameter information index is 0.046 and the index information error is given as. It implies that the dataset is viable for analysis because the value of error-index information is too low.

In table 2, the coefficient of determination shows that approximately 86% of the total variations were explained by the regressors.

Table 1. Final information measures

Objective Function Value	5.487485
Signal Entropy	6.139495
Noise Entropy	-0.65201
Normed Entropy (Signal)	0.953671
Normed Entropy (Noise)	0.999935
Parameter Information Index	0.046329
Error Information Index	0.000065

Table 2. Generalized maximum entropy summary of residual errors

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj. R. Sq.
Tillers	4	380	151842	395.4	19.8852	0.8628	0.8617

3.2. Generalized Maximum Entropy Variable Estimates

Table 3 provides the estimate of the generalized maximum entropy method of three parameters. It shows that the panicle is highly significant at 5% with a high influence on the estimated response (tiller) than the other independent variables. Also, Table gives a better estimate of the parameters than the robust regression in Table 4.1.

$\hat{y} = \text{Tiller}$

$X_1 = Panicle$

$X_2 = Plant\ Height$

$X_3 = Panicle\ length$

$\hat{y} = \hat{\beta}_{0(GME)} + \hat{\beta}_{1(GME)}X_1 + \hat{\beta}_{2(GME)}X_2 + \hat{\beta}_{3(GME)}X_3$

$\hat{y} = 8.7716 + 0.8774pan + 0.1298plant + 0.5649pan\_length$

Table 3. Generalized maximum entropy variable estimates

Variable	Estimate	Approx. Std. Err	t Value	Approx.
				Pr > |t|
Pan	0.877353	0.0193	45.46	<.0001
Plant_Ht	0.129803	0.0665	1.95	0.0518
Pan_Length	0.564892	0.4744	1.19	0.2345
Intercept	8.711554	12.2042	0.71	0.4758

3.3. Analysis of Variance

Table 4 shows the analysis of variance for the three independent variables which are; panicle, plant height and panicle length. It shows that they are highly significant at $(p < 0.05)$.

Table 4. Analysis of variance

Source	DF	Sum of	Mean	F Value	Pr > F
		Squares	Square		
Model	3	957573	319191	813.45	<.0001
Error	380	149110	392.3938		
Corrected Total	383	1106682			

4. Residuals

Figure 1 shows the residuals versus observation number. It shows that the deviation values are concentrated within the range -2 to +2. We see that most of the points lie close to zero and they are relatively homogeneous. Figure 1 also provides the regression model for the four variables and their relationship with the dependent variable. Figure 1 also shows an entropy regression, which is quite different from the generalized maximum entropy estimate. It shows the coefficient of determination is 0.8642 and root mean square error is 19.809. In table 5, we ran a robust regression to reduce outliers by finding a shrinkage estimator with the presence of two outliers.

$$\text{Tillers} = -14.324 + 0.8846\text{Pan} + 0.2527\text{Plant\_Ht} + 0.9473\text{Pan\_Length}$$

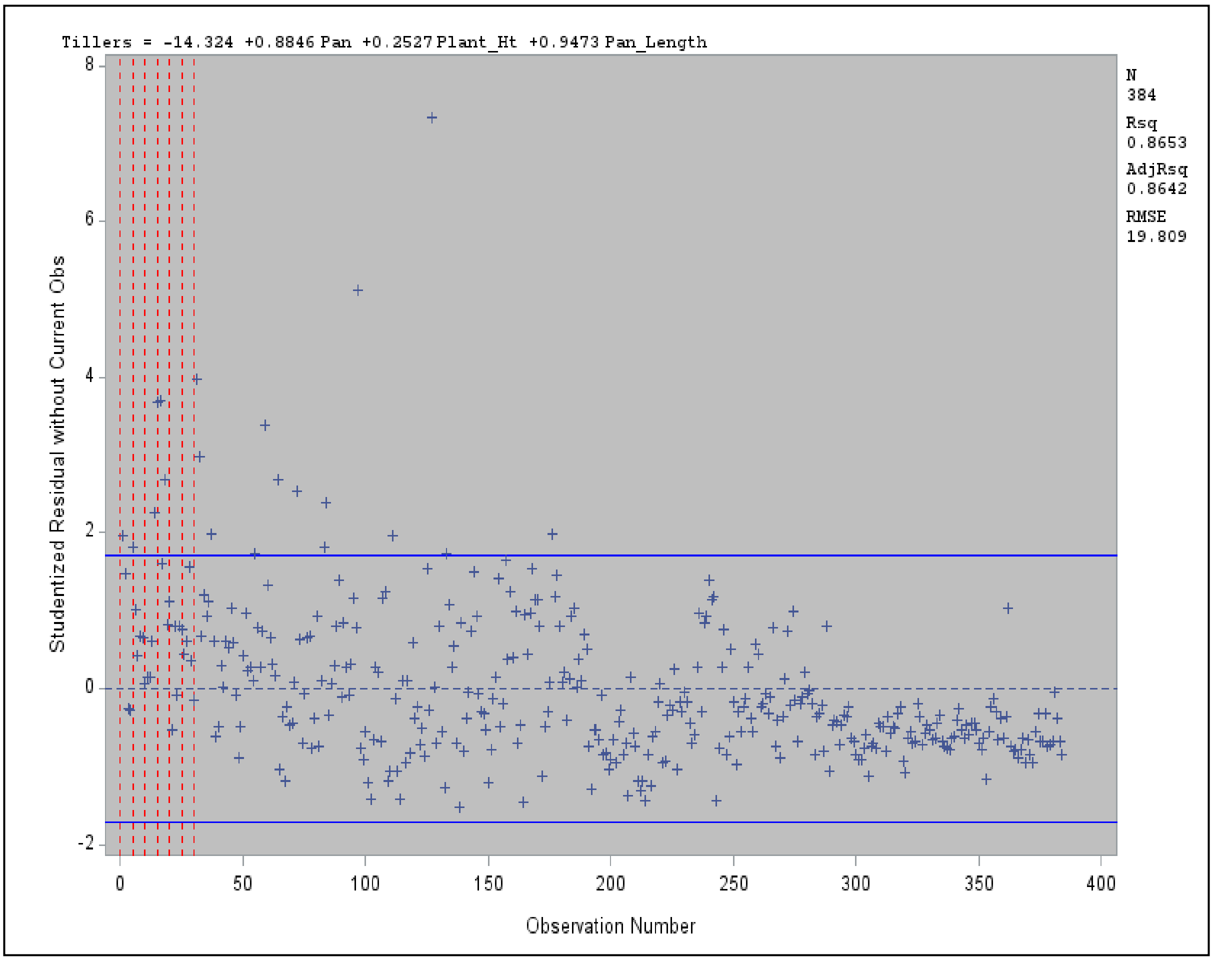

Figure 1. Showing residuals and observation points

$$\text{Tillers} = -11.2715 + 0.8986\text{Pan} + 0.1556\text{Plant\_Ht} + 1.08931\text{Pan\_Length}$$

Table 5. Robust regression analysis (Parameter estimates)

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-11.2715	10.2725	-31.4053	8.8623	1.2	0.2725
Pan	1	0.8986	0.0162	0.8667	0.9304	3060.11	<.0001
Plant_Ht	1	0.1556	0.056	0.0458	0.2654	7.72	0.0055
Pan_Length	1	1.0831	0.3993	0.3004	1.8658	7.36	0.0067

4.1. Fit Diagnostics for Dependent Variable (Weights)

Figure 2 shows that the regression is good as most points are not too far from the regression line with a good precision. The predicted value plot and the quantile plot depict a partially extreme value fit because they are almost linearly related. In figure 2, the diagnostic fit has no adequacy or validity threats to the dependent variable (Tiller).

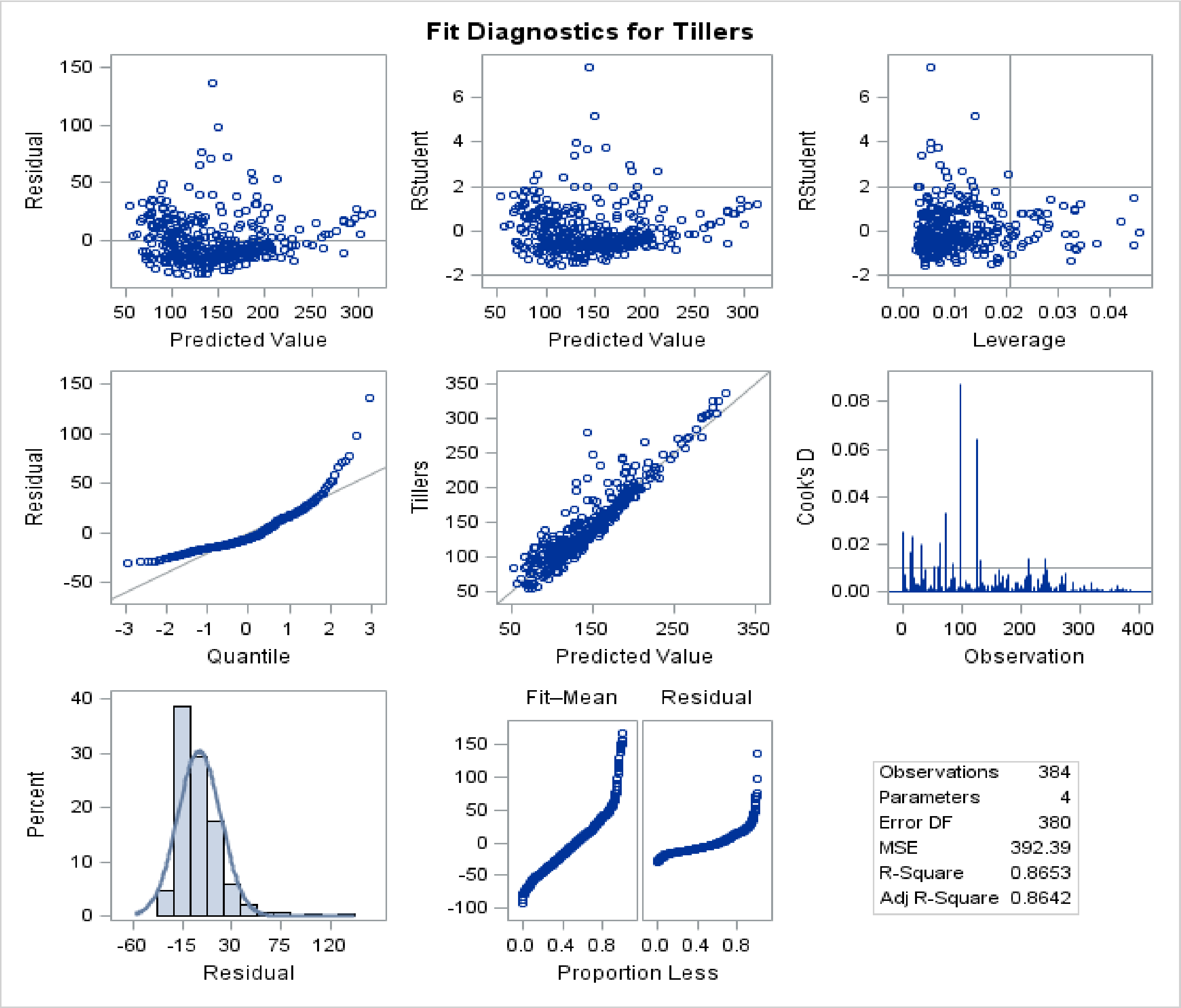

Figure 2. Diagnostics fit for Tillers

5. Conclusions

We have applied the generalized maximum entropy method for the estimation of four Agronomic variables. The result from the analysis shows that the GME-estimates differ from the robust regression-estimate in their intercepts. So, the maximum entropy provided a better estimate of these variables than the robust regression. In the model, panicle influence is high on the tiller than the other independent variables. We also found out that the panicle is highly significant at 5 percent from the analysis of variance and we detected two outliers from the residual analysis.

ACKNOWLEDGEMENTS

We appreciate the Sierra Leone Agricultural Research Institute for providing the research fields and part funding. We also thank Seibersdorf Laboratories for providing both radiation sources for rice varieties and the International Atomic Energy Agency (IAEA).

REFERENCES

[1] A Hald. 1952. “Statistical theory with engineering applications” John Wiley, New York.

[2] A Golan, G Judge. and D Miller 1996. Maximum entropy econometrics: Robust estimation with limited data, John Wiley & Sons.

[3] A Golan. and J. M Perloff, 2002. Comparison of maximum entropy and higher-order entropy estimators. J. Econ., vol. 107, pp. 195-211.

[4] C. E Shannon. (1948). A mathematical theory of communication. Bell. Syst. Tech. J., vol. 27, pp. 379-423. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.

[5] C Daniel., and F.S Wood., Fitting equations to data. Computer analysis of multifactor data. New York: John Wile/

[6] D. A Belsley, 1991. Conditioning diagnostics collinearity and weak data in regression. John Wiley: New York.

[7] E. T Jaynes. 1957. Information theory and statistical mechanics. Physics Review, 106, 620-630.

[8] E. Z Shen, and J. M Perlof. (2001). Maximum entropy and Bayesian approaches to the ratio problem. J. Appl. Econ., vol. 104, pp. 289313.

[9] F Pukelsheim 1994. The three sigma rule. American Statistician, 48, 88-91.

[10] H. K Mishra. 2004. Estimation under multicollinearity application of restricted Liu and maximum entropy estimators to the Portland Cement Dataset. MPRA Paper 1809, University Library of Munich, Germany.

[11] H Woods., H.H Steinor and H.R Starke. 1932. Effect of composition of Portland cement on Heat evolved During Hardening. Ind. Eng. Chem. Res., vol. 24, pp. 1207-1214.

[12] I Fraser, 2000. An application of maximum entropy estimation: The demand for meat in the United Kingdom. App. Econ., vol. 32, pp. 45-59.

[13] J. W Gorman and R.J Toman., 1966. Selection of variables for fitting equations to data. technometrics, vol. 8, pp. 27-51.

[14] J.N Kapur, and H.K Kesavan, (1992). Entropy optimization principles with applications. Academic Press, London.

[15] M. R Caputo. And Q Paris, 2008. Comparative statics of the generalized maximum entropy estimator of the general linear model. EJOR, vol. 185, pp. 195-203.

[16] R. C Campbell., and R. C Hill., A Monte Carlo study of the effect of design characteristics on the Inequality Restricted Maximum Entropy Estimator. Rev. App. Econ., vol. 1, pp. 53-84, 2005.

[17] R. C Campbell. and R.C Hill. 2006. Imposing parameter inequality restrictions using the principle of maximum entropy. J. Stat. Comput. Simul. vol. 76, pp. 985-1000.

[18] R Mittelhammer and S Cardell. 1997. On the consistency and asymptotic normality of the data constrained GME estimator of the GML. Working Paper, Washington State University, Pullman, WA.

[19] R Mittelhammer, G Judge, and D Miller, 2000. Econometric foundations, Cambridge University press.

[20] R Mittelhammer, S Cardell. and L. T Marsh, 2002. The Data Constrained GME Estimator of the GML: Asymptotic theory and inference. Working paper, Washington State University, Pullman, WA.

[21] S Kaçıranlar., S Sakallıolu., F Akdeniz., G.P.H Styan and H.J Werner, 1999. A new biased estimator in linear regression and a detailed analysis of the widely-analyzed data set on Portland cement (Sankhya., vol. 61, Series B, 3, pp. 443-459.

[22] X Wu, 2009. A weighted generalized maximum entropy estimator with a data-driven weight. Entropy, vol. 11, pp. 917-930.