Article

A New Entropy Optimization Model for Graduation of Data in Survival Analysis

Dayi He [1,2], Qi Huang [1,]* and Jianwei Gao [3]

[1] School of Humanities & Economics Management, Chinese University of Geosciences, Number 29, Xueyuan Rd., Haidian, Beijing 100083, China

[2] Lab of Resources & Environment Management, Chinese University of Geosciences, Number 29, Xueyuan Rd., Haidian, Beijing 100083, China

[3] School of Business Administration, North China Electronic Power University, Number 2, Beinong Rd., Changpin, Beijing 102206, China

* Author to whom correspondence should be addressed; E-Mail: huangqi@cugb.edu.cn; Tel.:+86-10-8232-2190; Fax: +86-10-8232-1783.

Abstract: Graduation of data is of great importance in survival analysis. Smoothness and goodness of fit are two fundamental requirements in graduation. Based on the instinctive defining expression for entropy in terms of a probability distribution, two optimization models based on the Maximum Entropy Principle (MaxEnt) and Minimum Cross Entropy Principle (MinCEnt) to estimate mortality probability distributions are presented. The results demonstrate that the two approaches achieve the two basic requirements of data graduating, smoothness and goodness of fit respectively. Then, in order to achieve a compromise between these requirements, a new entropy optimization model is proposed by defining a hybrid objective function combining both principles of MaxEnt and MinCEnt models linked by a given adjustment factor which reflects the preference of smoothness and goodness of fit in the data graduation. The proposed approach is feasible and more reasonable in data graduation when both smoothness and goodness of fit are concerned.

Keywords: entropy optimization; survival analysis; graduation of data

PACS Codes: 89.70.Cf; 02.50.Cw

1. Introduction

Survival analysis is an important topic in actuarial science. Survival analysis has a long history and there are many kinds of approaches. The common approach to estimate the survival distribution is the parametric one, with which a theoretical survival distribution is specified and the parameters involved are determined by certain methods. There are a lot of different methods available in literatures, such as maximum likelihood estimators [1–3] and Bayesian estimators [4–7], *etc*. To our best knowledge, the choice of theoretical survival distributions or other prior distributions is difficult and critical for the implementation of these kinds of methods. In this paper, we will discuss how to utilize an entropy optimization approach to estimate mortality distribution based on the instinctive relationship between entropy and probability distribution.

Information-theoretic entropy, presented by Shannon in 1948 [8] and the entropy optimization principle was proposed by Jaynes in 1957 [9,10] and Kullback *et al.* (1951, 1959) [11,12], have widened the application area of entropy and transformed it from a measure of information into a tool of statistics inference. Generally speaking, entropy optimization includes the Maximum Entropy Principle (MaxEnt) and Minimum Cross Entropy Principle (MinCEnt). MaxEnt estimates a probability distribution based only on the known information, without adding any other subjective information. MinCEnt is used to estimate a probability distribution which is closest to the prior one by minimizing the cross entropy between the estimated and the prior one.

Based on the instinctive defining expression for entropy in terms of a probability distribution, this paper will first review the idea of the application of entropy optimization rules in mortality distribution estimation. Then, in consideration of the goodness of fit and smoothness requirements in data graduation, a new approach will be proposed, which tries to combine the MaxEnt and MinCEnt methods to assure the degree of goodness of fit and smoothness of the estimation by use of an adjustment factor.

2. Review of Entropy Optimization Principles

There are two basic approaches to estimate probability distributions by using the concept of entropy: Maximum Entropy Principle (MaxEnt) and Minimum Cross Entropy Principle (MinCEnt). MaxEnt was proposed by Jaynes in 1957 [9,10]. He thought that given just some mean values, there are usually an infinity of compatible distributions. MaxEnt encourages us to select the distribution that maximizes the Shannon entropy measure and is simultaneously consistent with the mean value constraints. This is a natural extension of Laplace's famous *principle of insufficient reason*, which postulates that the uniform distribution is the most satisfactory representation of our knowledge when we know nothing about the random variant except that each probability is non-negative and that the sum of the probabilities is unity.

Let Θ be a discrete random variable on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where $\Omega = \{\theta_1, \theta_2, \cdots, \theta_n\}$ and $P(\Theta = \theta_i) = p_i$, $i = 1, 2, \cdots, n$ which is unknown and need to be estimated by some known information denoted by $g_i(\theta)(j = 1, 2, \cdots, m)$ such as mean value, variance, j^{th} moment, *etc*. Mathematically, MaxEnt can be described as the following optimization model:

$$\max \quad H(P) = -\sum_{i=1}^{n} p_i \ln p_i$$

$$\text{s.t.} \quad \begin{cases} \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} p_i g_j(\theta) = E_j, & j = 1, 2, \cdots, m \\ p_i \geqslant 0, & i = 1, 2, \cdots, n \end{cases} \tag{1}$$

where $H(P)$ is the entropy of a probability distribution P on Ω.

If there is a prior distribution of Θ, *i.e.*, $Q(\Theta = \theta_i) = q_i$, then MinCEnt can be used to get another estimated distribution which is statistically closest to the prior distribution under the same constraints as MaxEnt. MinCEnt can be modeled as:

$$\min \quad K(P, Q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

$$\text{s.t.} \quad \begin{cases} \sum_{i=1}^{n} p_i = 1 \\ \sum_{i=1}^{n} p_i g_j(\theta) = E_j, & j = 1, 2, \cdots, m \\ p_i \geqslant 0, & i = 1, 2, \cdots, n \end{cases} \tag{2}$$

where $K(P, Q)$ is the cross entropy between the probability distribution P and Q.

The two model can be solved by Lagrangian approach [13]. The solutions are:

$$p_i^{(\text{max})} = \exp\left[-\alpha_0 - \sum_{j=1}^{m} \alpha_j g_j(\theta)\right] \tag{3}$$

and:

$$p_i^{(\text{min})} = q_i \exp\left[-\beta_0 - \sum_{j=1}^{m} \beta_j g_j(\theta)\right] \tag{4}$$

where $\alpha_0, \alpha_1, \cdots, \alpha_m$ and $\beta_0, \beta_1, \cdots, \beta_m$ are Lagrangian multipliers for each model above.

3. Data Graduating with Entropy Optimization Principles

Assume that there is a living group whose original population is l_0. At time $x(x > 0)$, the living population is denoted as $l_x(x = 0, 1, \cdots, T)$. And we assume that $l_T = 0$. Let $d_x(x = 0, 1, \cdots, T)$ be the death population from time x to $x + 1$, then:

$$d_x = l_x - l_{x+1} \tag{5}$$

The value of d_x usually can be obtained from observation. In survival analysis, q_x is usually used to denote the mortality probability which means the probability for an individual not to live another year. The mortality probability is a basic parameter to construct a life table and has very important function in life insurance actuarial science. In real situations, it can be obtained from sample data and denoted as $\hat{q}_x$, which is:

$$\hat{q}_x = \frac{d_x}{l_x}$$

It is easy to find that $\sum_{x=0}^{T} \hat{q}_x \neq 1$. However, $\hat{q}_x$ is just an estimation of q_x which has to be graduated to approximate the real mortality probability as closely as possible. This process is called graduation of data in life insurance.

Furthermore, the estimation of q_x is based on the sample information, so we must utilize the information in sample fully and try our best to add as little extra information as possible. This may provide a reasonable foundation to use MaxEnt and MinCEnt. It should note that the mortality distribution does not meet the requirement of unity. Hence, in order to use MaxEnt and MinCEnt, let:

$$\tilde{q}_x = \frac{d_x}{l_0}, x = 0, 1, \cdots, T \tag{6}$$

and we define it as *death probability*, which is the ratio of the death population in different age groups to the original population. It is easy to find that $\tilde{q}_x$ can meet requirements of probability distribution, and it may be stated that:

$$q_0 = \tilde{q}_0, \qquad \tilde{d}_x = \tilde{l}_x - \tilde{l}_{x+1}, \qquad \hat{q}_x = \frac{\tilde{d}_x}{\tilde{l}_x}$$

Hence, $\hat{q}_x$, the estimation of q_x, can achieved once $\tilde{q}_x$ is determined.

Based on the sample information, the estimation of $\tilde{q}_x$ can be described as:

$$\max \quad H = -\sum_{x=1}^{T} \hat{q}_x \ln \hat{q}_x$$

$$\text{s.t.} \quad \begin{cases} \sum_{x=1}^{T} \hat{q}_x = 1 \\ \sum_{x=1}^{T} x\hat{q}_x = E_1 \\ \sum_{x=1}^{T} (x - E_1)^j \hat{q}_x = E_j, \quad j = 2, 3, \cdots, k \\ \tilde{q}_x > 0, \qquad\qquad x = 1, 2, \cdots, T \end{cases} \tag{7}$$

where $\hat{q}_x$ is the estimation of $\tilde{q}_x$, E_1 is the mean value of sample data and E_j is the jth moment.

On the other hand, $\tilde{q}_x$ can be viewed as a prior distribution of death probability, then a MinCEnt model can be established to estimation of $\tilde{q}_x$ as:

$$\min \quad K = -\sum_{x=1}^{T} \hat{q}_x \ln \frac{\hat{q}_x}{\tilde{q}_x}$$

$$\text{s.t.} \quad \begin{cases} \sum_{x=1}^{T} \hat{q}_x = 1 \\ \sum_{x=1}^{T} x\hat{q}_x = E_1 \\ \sum_{x=1}^{T} (x - E_1)^j \hat{q}_x = E_j, \quad j = 2, 3, \cdots, k \\ \tilde{q}_x > 0, \qquad\qquad x = 1, 2, \cdots, T \end{cases} \tag{8}$$

The solution of Equations (7) and (8) can be achieved by using Lagrangian approaches [13]. The results are:

$$\hat{q}_x^{(\max)} = \exp\left[-\sum_{j=0}^{k} \alpha_j (x - E_1)^j\right] \tag{9}$$

and:

$$\hat{q}_x^{(\min)} = \tilde{q}_x \exp\left[-\sum_{j=0}^{k} \beta_j (x - E_1)^j\right] \tag{10}$$

where $\alpha_j, \beta_j (j = 0, 1, \cdots, k)$ are Lagrangian multipliers.

To clarify the feasibility and properties of the above estimation methods, the above models will be applied to estimate the mortality distribution on experimental data taken from Ananda *et al*. [5].

Example: The following data (Table 1) are death times for 208 mice, which were exposed to gamma radiation. The data are divided into 14 groups by the time interval given in [5] and then we calculate the mortality distribution by the MaxEnt and MinCEnt model.

Table1. Experimental data.

Time Interval(x)	Death Population (d_x)	Death Probability ($\tilde{q}_x$)
1	3	0.0144
2	3	0.0144
3	6	0.0288
4	6	0.0288
5	16	0.0769
6	14	0.0673
7	25	0.1202
8	20	0.0962
9	32	0.1530
10	25	0.1202
11	27	0.1298
12	13	0.0625
13	11	0.0529
14	7	0.0337

With the MaxEnt and MinCEnt approach, Table 2 shows the data graduation results under different moment constraints (up to 5th), and they are plotted in Figures 1 and 2 too.

Table 2. Results of MaxEnt and MinCEnt.

Experiment Data		Results of MaxEnt(up to E_j)					Results of MinCEnt(up to E_j)				
x	$\tilde{q}_x$	E_1	E_2	E_3	E_4	E_5	E_1	E_2	E_3	E_4	E_5
1	0.0144	0.0447	0.0074	0.0115	0.0131	0.0126	0.0144	0.0144	0.0144	0.0144	0.0144
2	0.0144	0.0478	0.0147	0.0177	0.0177	0.0180	0.0144	0.0144	0.0144	0.0144	0.0144
3	0.0288	0.0511	0.0264	0.0271	0.0258	0.0263	0.0288	0.0288	0.0288	0.0288	0.0288
4	0.0288	0.0546	0.0434	0.0407	0.0386	0.0389	0.0288	0.0288	0.0288	0.0288	0.0288
5	0.0769	0.0583	0.0648	0.0587	0.0570	0.0567	0.0769	0.0769	0.0770	0.0769	0.0770
6	0.0673	0.0624	0.0882	0.0806	0.0805	0.0797	0.0673	0.0673	0.0674	0.0674	0.0674

Table 2. *Cont.*

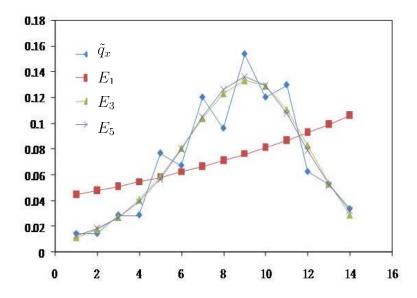| Experiment Data		Results of MaxEnt(up to E_j)					Results of MinCEnt(up to E_j)				
x	$\tilde{q}_x$	E_1	E_2	E_3	E_4	E_5	E_1	E_2	E_3	E_4	E_5
7	0.1202	0.0667	0.1094	0.1036	0.1058	0.1050	0.1202	0.1202	0.1203	0.1203	0.1204
8	0.0962	0.0712	0.1237	0.1230	0.1264	0.1265	0.0962	0.0962	0.0962	0.0963	0.0963
9	0.1538	0.0762	0.1274	0.1329	0.1353	0.1363	0.1538	0.1538	0.1538	0.1538	0.1538
10	0.1202	0.0814	0.1195	0.1288	0.1284	0.1293	0.1202	0.1202	0.1202	0.1202	0.1201
11	0.1298	0.0870	0.1022	0.1105	0.1076	0.1076	0.1298	0.1298	0.1297	0.1297	0.1297
12	0.0625	0.0930	0.0796	0.0827	0.0797	0.0789	0.0625	0.0625	0.0625	0.0624	0.0625
13	0.0529	0.0994	0.0565	0.0532	0.0527	0.0520	0.0529	0.0529	0.0529	0.0529	0.0529
14	0.0337	0.1063	0.0366	0.0290	0.0315	0.0321	0.0337	0.0337	0.0337	0.0337	0.0337
χ^2		0.3237	0.0432	0.0342	0.0334	0.0334	0.0000	0.0000	0.0000	0.0000	0.0000
$[\Delta^4 \hat{q}_x]^2$		0.0000	0.0000	0.0001	0.0001	0.0001	0.6184	0.6184	0.6184	0.6184	0.6184

Figure 1. Results of MaxEnt.
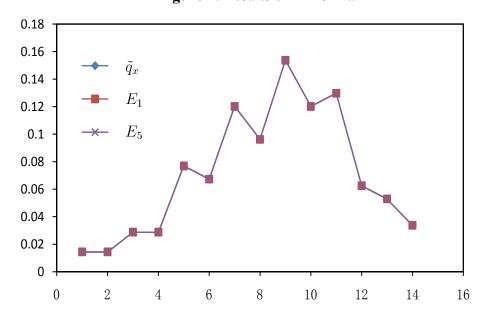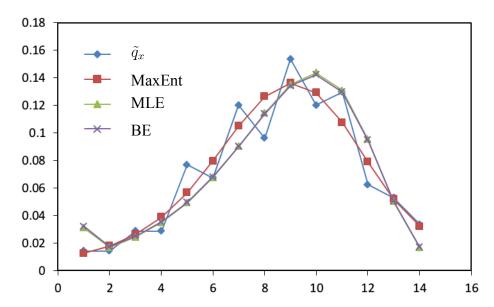

Figure 2. Results of MinCEnt.

Table 3 and Figure 3 are the comparison of results of MaxEnt (up to 5th moment) with results of maximum likelihood estimation (MLE) and Bayesian estimation (BE) method from [5].

Table3. Results of Different Graduation Approaches.

x	$\tilde{q}_x$	MaxEnt	MLE	BE
1	0.0144	0.0126	0.0311	0.0321
2	0.0144	0.0180	0.0168	0.0173
3	0.0288	0.0263	0.0244	0.0249
4	0.0288	0.0389	0.0349	0.0354
5	0.0769	0.0567	0.0491	0.0496
6	0.0673	0.0797	0.0676	0.0680
7	0.1202	0.1050	0.0901	0.0902
8	0.0962	0.1265	0.1144	0.1140
9	0.1538	0.1363	0.1351	0.1342
10	0.1202	0.1293	0.1437	0.1423
11	0.1298	0.1076	0.1310	0.1297
12	0.0625	0.0789	0.0953	0.0948
13	0.0529	0.0520	0.0501	0.0505
14	0.0337	0.0321	0.0165	0.0170
χ^2		0.0334	0.0756	0.0737
$[\Delta^4 \tilde{q}_x]^2$		0.0001	0.0008	0.0007

Figure 3. Results of Different Graduation Approaches.

From above results, it can be found that:

(1) From Table 3, results of MaxEnt have smaller χ^2 value(a measure of goodness of fit to original data and calculated by Euclidean distance) and $[\Delta^4 \hat{q}_x]^2$ (a measure of smoothness by using 4th-moment of difference) than the results of MLE and BE approach. Hence, MaxEnt method of data graduating can be thought as a better method.

(2) From Table 2, results of the MinCEnt approach are the same as the prior distribution, and the reason is that $\tilde{q}_x$ and E_j are calculated from the experimental data. If $\tilde{q}_x$ or E_j is given by other information, the results will be different from the prior distribution. However, from this extreme situation we can find that MinCEnt focuses on the goodness of fit in data graduation.

(3) Comparing with the MinCEnt approach, the MaxEnt approach is better from the viewpoint of the smoothness of data graduation and is worse from the point of view of goodness of fit.

4. Data Graduating by Combining MaxEnt and MinCEnt

Smoothness and goodness of fit always are the most important consideration in graduation of data though many techniques have been developed. For example, the most widely used method until now, especially by North American actuaries for the construction of life tables, is the Whittaker-Henderson method of graduation. This method originates in the work of Bohlmann (1899) [14] and Whittaker (1923) [15], and contributions to the theory were made by Henderson (1924, 1925) [16,17], and others. The Whittaker-Henderson method gives the graduated values by minimizing the quantity:

$$M = F + hS = \sum_{x=1}^{n} \omega_x (v_x - u_x)^2 + h \sum_{x=1}^{n-z} [\Delta^z v_x]^2 \tag{11}$$

where F and S are the weighted measure of goodness of fit to the original data and smoothness receptively. u_x is the prior of mortality, v_x is the estimated mortality, *i.e.*, result of data graduation, and $\Delta^z v_x$ is the zth difference of v_x (usually $z = 3, 4$ or higher). ω_x is a weight coefficient and h is a positive adjustment factor between goodness of fit and smoothness. This method is widely used and has become a basic logic in graduation of data and many other approaches currently follow this lead. Generally speaking, the characteristics of different data graduation methods may lie on two sides, putting different emphasis on the goodness of fit and smoothness, and on how to measure smoothness and goodness of fit. Therefore, graduation of data can be looked as a bi-objective question. On one hand, the graduation results should be smooth and on the other hand they should be close to the original data. From the above section, it has shown that results of MaxEnt data graduation is smoother while those of MinCEnt is closer to the original data, so we propose a new approach of data graduation which can combine the both methods as in the following model:

$$\min \quad G = \mu \sum_{x=1}^{T} \hat{q}_x \ln \hat{q}_x + (1 - \mu) \sum_{x=1}^{T} \hat{q}_x \ln \frac{\hat{q}_x}{\tilde{q}_x}$$

$$\text{s.t.} \quad \begin{cases} \displaystyle\sum_{x=1}^{T} \hat{q}_x = 1 \\ \displaystyle\sum_{x=1}^{T} x \hat{q}_x = E_1 \\ \displaystyle\sum_{x=1}^{T} (x - E_1)^j \hat{q}_x = E_j, \quad j = 2, 3, \cdots, k \\ \tilde{q}_x > 0, \qquad\qquad\quad x = 1, 2, \cdots, T \end{cases} \tag{12}$$

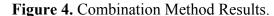where $0 \leqslant \mu \leqslant 1$ is a given adjustment factor between smoothness and goodness of fit. When $\mu = 0$, it is the MinCEnt approach, and when $\mu = 1$ it is the MaxEnt approach. Because Equations (7) and (8) are solvable, it is easy to conclude that Equation (12) is solvable too.
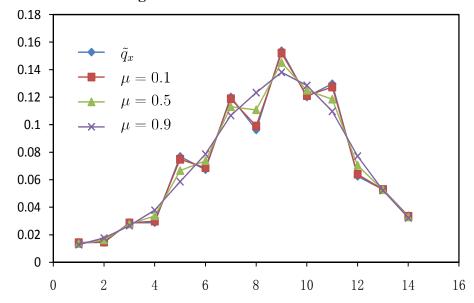
In the above model, MaxEnt is proposed as a measure of smoothness and MinCEnt is proposed as a measure of goodness of fit. These two measures are integrated with a linear coefficient μ to reflect different weights on smoothness and goodness of fit. The reason to adopt a convex combination of smoothness and goodness of fit is to assure convexity of the objective function G and solvability of the proposed model. In reality, how to decide the appropriate weight between smoothness and goodness of fit is a highly controversial topic. We propose that the value of μ can be determined as h determined in Equation (11), which is usually determined by experimental approaches.

Based on the data of above Example, we calculate the estimated mortality distribution by this model. Table 4 and Figure 4 are results of this approach when taking 5th moment as constraints.

Table 4. Combination Method Results.

Experimental Data		Combination Method				
x	$\tilde{q}_x$	$\mu = 0.1$	$\mu = 0.3$	$\mu = 0.5$	$\mu = 0.7$	$\mu = 0.9$
1	0.0144	0.0142	0.0139	0.0135	0.0132	0.0128
2	0.0144	0.0147	0.0154	0.0162	0.0169	0.0176
3	0.0288	0.0286	0.0282	0.0277	0.0272	0.0266
4	0.0288	0.0298	0.0317	0.0337	0.0358	0.0378
5	0.0769	0.0749	0.0707	0.0666	0.0626	0.0586
6	0.0673	0.0687	0.0713	0.0738	0.0762	0.0786
7	0.1202	0.1189	0.116	0.113	0.1099	0.1067
8	0.0962	0.0991	0.1048	0.1108	0.1169	0.1233
9	0.1538	0.1521	0.1487	0.1452	0.1416	0.1381
10	0.1202	0.1211	0.1231	0.125	0.1268	0.1285
11	0.1298	0.1274	0.123	0.1185	0.1141	0.1097
12	0.0625	0.064	0.0672	0.0705	0.0738	0.0772
13	0.0529	0.0529	0.0528	0.0527	0.0525	0.0522
14	0.0337	0.0335	0.0332	0.0329	0.0326	0.0323

Figure 4. Combination Method Results.

From the above results, it can be concluded that the proposed method is feasible and the adjustment factor plays an important role in trading off smoothness and goodness of fit, which provides flexibility in graduation of data.

4. Conclusions

In this paper, based on the instinctive defining expression for entropy in terms of a probability distribution, the methods of application of entropy optimization in survival analysis were discussed. It was found that the results of the MaxEnt model focus on the smoothness of data graduation, while results of the MinCEnt model focus on the goodness of fit to the original data, so in consideration of the requirements of smoothness and goodness of fit in data graduation, a new approach was proposed to combine the results of MaxEnt and MinCEnt, which could provide a trade-off between smoothness and goodness of fit in data graduation by use of a given adjustment factor.

Acknowledgments

References

1. Broffitt, J.D. Maximum likelihood alternatives to actuarial estimators of mortality rates. *Trans. Soc. Actuar.* **1984**, *36*, 77–142.
2. Zhou, J.; Liu, J.; Li, Y. *The Theory of Constructing Life Table*; Nankai University Press: Tianjin, China, 2001.
3. Dennis, T.H.; Fellingham, G.W. Likelihood methods for combining tables of data. *Scand. Actuar. J.* **2000**, *2*, 89–101.
4. Singh, A.K.; Ananda, M.A.; Dalpatadu, R. Bayesian estimation of tabular survival models from complete samples. *Actuar. Res. Clear. House* **1993**, *1*, 335–342.
5. Ananda, M.M.; Dalpatadu, R.J.; Singh, A.K. Estimating parameters of the force of mortality in actuarial studies. *Actuar. Res. Clear. House* **1993**, *1*, 129–141.
6. Haastrup, S. Comparison of some Bayesian analysis of heterogeneity in group life insurance. *Scand. Actuar. J.* **2000**, *1*, 2–16.
7. Nielson, A.; Lewy, P. Comparison of the frequentist properties of Bayes and the maximum likelihood estimators in an age structured fish stock assessment model. *Can. J. Fish. Aquat. Sci.* **2002**, *59*, 136–143.
8. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
9. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* 1957, 106, 620–630.
10. Jaynes, E.T. Information theory and statistical mechanics II. *Phys. Rev.* 1957, 108, 171–190.
11. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
12. Kullback, S. *Information Theory and Statistics*; John Wiley and Sons: New York, NY, USA, 1959.

13. Kapur, J.N.; Kesavan, H.K. *Entropy optimization Principles with Applications*; Academic Press Inc.: San Diego, CA, USA, 1992.

14. Bohlmann, G. Ein Ausgleichungs Problem. In *Nachrichten von der Königl Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*; Horstmann, L., Ed.; Commissionsverlag der Dieterich'schen Universitätsbuchhandlung: Göttingen, Germany, 1899; pp. 260–271.

15. Whittaker, E.T. On a new method of graduation. *Proc. Edinb. Math. Soc.* **1923**, *41*, 63–75.

16. Henderson, R. A new method of graduation. *Trans. Actuar. Soc. Am.* **1924**, *25*, 29–53.

17. Henderson, R. Further remarks on graduation. *Trans. Actuar. Soc. Am.* **1925**, *26*, 52–74.

18. Alicja, S.N.; William, F.S. An extension of the Whittaker-Henderson method of graduation. *Scand. Actuar. J.* **2012**, *1*, 70–79.