Paper 120,CCG Annual Report 12, 2010 (©2010)

Probability Estimation with Maximum Entropy Principle

Yupeng Li and Clayton V. Deutsch

The principle of Maximum Entropy is a powerful and versatile tool for inferring a probability distribution from constraints that do not completely characterize the distribution. The principle of Minimum Relative Entropy which is a more general form of Maximum Entropy method has all the important attributes of the Maximum Entropy method with the advantage of more easily integration of the prior probability distribution. The maximum entropy methods have been successfully explored in many disciplines. While used in discrete multivariate probability distribution estimation, there are some challenges with the traditional Lagrange multiplier approach to Maximum Entropy and Minimum Relative Entropy. In this paper, the Iterative Scaling based on Minimum Relative Entropy is used in discrete multivariate probability estimation which makes the Minimum Relative Entropy principle approach successfully implementation in discrete multivariate probability estimation. The principle of Maximum Entropy and Minimum Relative Entropy are introduced. Then, the challenges in discrete Multivariate Probability estimation are illustrated with a small discrete probability estimation example. A solution based on iterative scaling is introduced and explained with two numerical application examples.

1 Maximum Entropy Principle

The principle of maximum entropy (ME) is based on the premise that *when estimating the probability distribution, the best estimation results will keep the largest remaining uncertainty (the maximum entropy) consistent with all the known constraints. In that way, no more additional assumptions or biases were introduced in the estimation* [1, 2].

Generally, assuming any desired multivariate probability is $\mathbf{p} : \{p_\ell, \ell = 1, \cdots, N\}$, the objective function will be satisfied from some constraints:

$$\texttt{Maximum:} \qquad H(\mathbf{p}) = -\sum_{\ell=1}^{N} p_\ell Log(p_\ell) \tag{1}$$

subject to:

$$\sum p_\ell = 1 \tag{2}$$

$$\sum a_{m\ell} p_\ell = b_m; m = 1, 2, \cdots, M \tag{3}$$

Where b_m are any order of marginal for the desired discrete multivariate probability.

Theorem 1 *Only under the normal constraints, the uniform distribution will be the maximum entropy solution.*

Proof The well-known solution to the problem of optimizing a function subject to constraints is the method of Lagrange multipliers[3]. The first step is to form a new entropy function $L(p_\ell, \lambda)$ as defined below:

$$L = -\sum_{\ell} p_\ell \log(p_\ell) + \lambda\big(\sum_{\ell} p_\ell - 1\big) \tag{4}$$

The second step is equating the derivate of (4) to zero with respect to each variables $p_\ell, \ell = 1, \cdots, N$ and λ. The results will be an equation set:

$$\frac{\partial L}{\partial p_\ell} = -1 - \log(p_\ell) + \lambda = 0 \tag{5}$$

$$\frac{\partial L}{\partial \lambda} = \sum_\ell (p_\ell) - 1 = 0 \tag{6}$$

From equation (5), it is obtained that $p_\ell = exp(\lambda - 1)$. This is independent of ℓ, Thus, all the probabilities p_ℓ should be equal and sum to 1. Then the uniform distribution $p_\ell = 1/N$ is the ME estimation.

When the full constraints are enforced to the objective function. Using the method of Lagrange multipliers, the new objective function $L(p, \lambda, \lambda_m)$ will be defined as:

$$L(p_\ell, \lambda, \lambda_m) = -\sum_\ell p_\ell \log(p_\ell) + \lambda\big(1 - \sum_\ell p_\ell\big) + \lambda_m\big(b_m - \sum_\ell a_{m\ell} p_\ell\big) \tag{7}$$

Then equate the derivative of equation (7) to zero with respect to each of the variables $p_\ell, \ell = 1, \cdots, N$ and $\lambda, \lambda_m; m = 1, \cdots, M$ that is:

$$\frac{\partial L}{\partial p_\ell} = -Log p_\ell - 1 - \lambda - \sum_{m=1}^{M} \lambda_m a_{m\ell} = 0 \tag{8}$$

$$\frac{\partial L}{\partial \lambda} = 1 - \sum_\ell p_\ell = 0 \tag{9}$$

$$\frac{\partial L}{\partial \lambda_m} = b_m - \sum_\ell a_{m\ell} p_\ell = 0 \tag{10}$$

Then from equation (8), it is obtained that

$$p_\ell = exp\Big(- \lambda_0 - \sum_{m=1}^{M} \lambda_m a_{m\ell}\Big) \quad (\lambda_0 = \lambda + 1) \tag{11}$$

In information theory and statistical mechanics, the preferred form is written as:

$$p_\ell = \frac{1}{\Lambda} exp(- \sum_{m=1}^{M} \lambda_m a_{m\ell}) \tag{12}$$

Where Λ is called partition function which is a function between λ_0 and all the other λ_m and written as:

$$\Lambda = exp(\lambda_0) = \sum_{\ell=1}^{N} \pi_\ell exp(- \sum_{m=1}^{M} \lambda_m a_{m\ell}) \tag{13}$$

The probability law in (11) is also named as the Gibbs distribution. The Lagrange multiplier λ_0 is called the potential equation which has the property that:

$$\frac{\partial \lambda_0}{\lambda_m} = -b_m, \quad m = 1, \cdots, M \tag{14}$$

Theoretically, from equation (14), all the needed $\lambda_m; m = 1, \cdots, M$ can be solved from the m equations. However, solving the m set of coupled implicit nonlinear equation is a difficult task in practice.

2 Minimum relative entropy principle

In probability and information theory, the relative entropy which is also known as Kullback-Leibler divergence, information divergence, information gain is an asymmetric measure of the difference between two probability distributions [4, 5, 6, 7]. Consider two sets of discrete probability $(p_1, \cdots, p_N)$ and $(\pi_1, \cdots, \pi_N)$. The relative entropy between them which is also a measure of the difference of the information contained in them is defined as:

$$J[p \parallel \pi] = \sum_{\ell} p_\ell \log \frac{p_\ell}{\pi_\ell} \tag{15}$$

It is well known that:

$$J[p \parallel \pi] \geq 0 \tag{16}$$

$$J[p \parallel \pi] = 0 \quad \text{if only if} \quad p = \pi \tag{17}$$

The Principle of Minimum Relative Entropy(MRE) is that *given new facts, a new distribution p should be chosen which is as hard to discriminate from the original distribution π as possible so that the new data produce as small an information gain $J(p||\pi)$ as possible, thus no more bias except satisfying the constraints are introduced.* [8] It is written as:

$$\texttt{Minimize:} \quad J(p \parallel \pi) = \sum_{\ell} p_\ell \log \frac{p_\ell}{\pi_\ell} \tag{18}$$

subject to:

$$\sum p_\ell = 1 \tag{19}$$

$$\sum a_{m\ell} p_\ell = b_m; m = 1, 2, \cdots, M \tag{20}$$

where $a_{m\ell}$ is the marginal construction matrix. The constraint b_m will be the marginal probability and it will satisfy $\sum_{m=1}^{M} b_m = 1$.

Theorem 2 *Maximum Entropy is equivalent to Minimum Relative Entropy principle between P and the uniform distribution $\mathfrak{U}$.*

Proof Given the uniform probability distribution $\mathfrak{U} = 1/N$, then Minimum KL divergence results will be :

$$\begin{aligned} \texttt{Minimize:} \quad J(P \parallel \mathfrak{U}) &= \sum_{\ell=1}^{N} p_\ell \log(p_\ell N) \\ &= \sum_{\ell=1}^{n} p_\ell \log p_\ell + N \log N \\ &= -H(P) + (\text{constant}) \end{aligned} \tag{21}$$

As shown in Equation (21), the results from minimizing $J(p \parallel \pi)$ will bring maximum entropy results to $H(p)$.

Thus, the ME approach is a special case of the MRE. However, the MRE formulation is more general and offers greater flexibility for the null-hypothesis function π can represent any probability function.

2.1 Solving MRE by Lagrange multiplier

The same Lagrange multipliers approach is used in solving the solution. The first step is to form a new entropy function $L(p_\ell, \lambda, \lambda_m)$ as defined below:

$$L(p_\ell, \lambda, \lambda_m) = -\sum_{\ell=1}^{N} p_\ell \log \frac{p_\ell}{\pi_\ell} + (\lambda_0 - 1)\big(\sum_{\ell=1}^{N} p_\ell - 1\big) + \lambda_m\big(\sum_{\ell=1}^{N} a_{m\ell} p_\ell - b_m\big) \tag{22}$$

Then equate the derivative of equation (22) to zero with respect to each desired probabilities $p_\ell, \ell = 1, \cdots, N$ and all the Lagrange multipliers, that is:

$$\frac{\partial L}{\partial p_\ell} = -\log \frac{p_\ell}{\pi_\ell} - 1 - \lambda_0 + 1 - \sum_{m=1}^{M} \lambda_m a_{m\ell} = 0 \tag{23}$$

$$\frac{\partial L}{\partial \lambda_0} = \sum_{\ell} p_\ell - 1 = 0 \tag{24}$$

$$\frac{\partial L}{\partial \lambda_m} = b_m - \sum_{\ell} a_{m\ell} p_\ell = 0 \tag{25}$$

From equation (23),(24) and (25), the final estimated probability could be expressed as:

$$p_\ell = \pi_\ell exp\Big(- \lambda_0 - \sum_{m=1}^{M} \lambda_m a_{m\ell}\Big) \tag{26}$$

Or

$$p_\ell = \frac{1}{\Lambda} exp(- \sum_{m=1}^{M} \lambda_m a_{m\ell}) \tag{27}$$

Where equation (27) is the expression that used most often in information theory. Same as the ME solution, the expression Λ is called partition function which is a function of all the Lagrange multipliers as:

$$\Lambda = exp(\lambda_0) = \sum_{\ell=1}^{N} \pi_\ell exp(- \sum_{m=1}^{M} \lambda_m a_{m\ell}) \tag{28}$$

Where λ_0 is called the potential equation. Same as the ME, from the potential equation, taking the derivative according to all the other Lagrange multipliers, the result will also be a coupled nonlinear equation set. It could be solved with some numerical approach for very simple problems.

2.2 Solving MRE using iterative scaling

From previous sections, it is shown that the maximum entropy solution will be a coupled nonlinear equation set. It could be possible solved with some numerical approach. But until now, no clear and transparent solution have been given. In this section, one kind of iterative scaling solution to the MRE is adopted[4, 9].

The MRE from iterative scaling is said that: Given the constraints in equation (20), there exists an unique probability distribution $\hat{P}_\ell$ which satisfies them and is the limit of the iterative sequence $\{\mathbf{p}^{(\delta)}; \delta = 0, 1, 2, \cdots\}$ defined by

$$\hat{p}_\ell^{(0)} = \pi_\ell$$

$$\hat{p}_\ell^{(\delta+1)} = \hat{p}_\ell^{(\delta)} \mu \prod_{m=1}^{M} \left[\frac{b_m}{\hat{b}_m^{(\delta)}}\right]^{a_{m\ell}} \quad \delta = 0, 1, 2, \cdots \tag{29}$$

where $\hat{b}_m^{(\delta)} = \sum a_{m\ell}\hat{p}_\ell^{(\delta)}$, and μ is used to do normalization.

The proof of unique and convergence of this iterative process in Equation (29) is given by Kullback and Khairat in 1966[7]. This method is named as **Iterative Scaling(IS)** which is continued studied extensively in mathematics and statistic researches[10, 11, 12, 13, 14].

More generally, consider R sets of constraints each of them is in the form of (20), Let the r^{th} set constraint be written as:

$$\sum_\ell a_{m\ell}^r p_\ell = b_m^r, \quad r = 1, 2, \cdots, R; \quad m = 1, \cdots, M \tag{30}$$

where $\sum_{m=1}^M b_m^r = 1$. In the multivariate probability estimation, this r^{th} set constraints could be any order lower marginal probability. Provided that the constraints in (30) are consistent to each other, there exists a unique positive probability distribution $\mathbf{p}$ which satisfies them and is of the form:

$$p_\ell = \pi_\ell \mu \prod_{r=1}^{R} \prod_{m=1}^{M} (\mu_m^r)^{a_{m\ell}^r} \tag{31}$$

which means that $\mathbf{p}$ can be obtained as the limit of a "cyclical" iterative scaling process[9, 10]. The starting probability of the iteration π can take the uniform distribution which was simplest and most natural choice (Darroch and Ratcliff ,1972)[11]. A more reasonable and practical choice of π is assuming that all variables are independent. that is the initial estimation will be $\pi = \prod P(\mathbf{u}_\alpha)$. Thus, the estimated multivariate represent a "generalized" independent distribution subject to the linear constraints(lower-marginal distributions).

Here a simple discrete probability $P = \{p_1, p_2, p_3, p_4, p_5\}$ is used to show one iterative scaling process. Let the linear constraints are

$$\sum_{\ell=1}^{5} a_{m\ell} p_i = b_m, \quad m = 1, 2, 3, 4; \quad \ell = 1, 2, 3, 4, 5 \tag{32}$$

Given an marginal construction matrix $a_{m\ell}$, the linear constraints in Equation (32) written in traditional matrix form will be:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix} \tag{33}$$

Assuming the initial probability distribution is $\{p_1^{(0)}, p_2^{(0)}, p_3^{(0)}, p_4^{(0)}, p_5^{(0)}\}$, the first iterative scaling process

will be proceed as:

$$p_1^{(1)} = p_1^{(0)}(\frac{h_1}{h_1^{(0)}})^1(\frac{h_2}{h_2^{(0)}})^0(\frac{h_3}{h_3^{(0)}})^0(\frac{h_4}{h_4^{(0)}})^1 = p_1^{(0)}\prod_{r=1}^{4}(\frac{h_r}{h_r^{(0)}})^{a_{r1}}$$

$$p_2^{(1)} = p_2^{(0)}(\frac{h_1}{h_1^{(0)}})^1(\frac{h_2}{h_2^{(0)}})^1(\frac{h_3}{h_3^{(0)}})^0(\frac{h_4}{h_4^{(0)}})^1 = p_2^{(0)}\prod_{r=1}^{4}(\frac{h_r}{h_r^{(0)}})^{a_{r2}}$$

$$p_3^{(1)} = p_3^{(0)}(\frac{h_1}{h_1^{(0)}})^0(\frac{h_2}{h_2^{(0)}})^1(\frac{h_3}{h_3^{(0)}})^0(\frac{h_4}{h_4^{(0)}})^0 = p_3^{(0)}\prod_{r=1}^{4}(\frac{h_r}{h_r^{(0)}})^{a_{r3}}$$

$$p_4^{(1)} = p_4^{(0)}(\frac{h_1}{h_1^{(0)}})^0(\frac{h_2}{h_2^{(0)}})^0(\frac{h_3}{h_3^{(0)}})^1(\frac{h_4}{h_4^{(0)}})^1 = p_4^{(0)}\prod_{r=1}^{4}(\frac{h_r}{h_r^{(0)}})^{a_{r4}}$$

$$p_5^{(1)} = p_5^{(0)}(\frac{h_1}{h_1^{(0)}})^1(\frac{h_2}{h_2^{(0)}})^1(\frac{h_3}{h_3^{(0)}})^1(\frac{h_4}{h_4^{(0)}})^0 = p_5^{(0)}\prod_{r=1}^{4}(\frac{h_r}{h_r^{(0)}})^{a_{r5}}$$

so after the first iteration the probability will be:

$$p_i^{(1)} = p_i^{(0)}\mu\prod_{r=1}^{4}[\frac{h_r}{h_r^{(0)}}]^{a_{ri}} \qquad i = 1, \cdots, 5 \tag{34}$$

where μ is used to normalization to make it a probability. Using the same iterative scaling process the iterated sequence $\{\mathbf{p^{(n)}}; n = 0, 1, 2, \cdots\}$ can be obtained, the limit of this sequence would be the solution satisfying the linear constraints.

3 Numerical example of iterative scaling

The die problem

The die problem serves as an illustration and of different solution efforts from iterative scaling and Lagrange multiplier. This problem was originally proposed by Jaynes as an example to show the ME principle for undetermined problem[3]. It will show there is no difference in the final results from using Lagrange multiplier and iterative scaling approach. However, the procedure of IS is more straightforward than the Lagrange approach.

Consider a die of six faces is tossed for $T(T \to +\infty)$ times. One is told that the average number of spots up was not 3.5 as we might expected from an "honest" die but 4.5. Only given this information, and nothing else, what probability should one assign to i spots on the next toss?

From maximum entropy approach, the solution could be proceed as following procedures. The constraints to the entropy equation would be:

$$\sum_{i=1}^{6} i \cdot p_i = 4.5 \tag{35}$$

$$\sum_{i=1}^{6} p_i = 1 \tag{36}$$

the new objective function with Lagrange multipliers would be:

$$L = -\sum_{i=1}^{6} p_i \log p_i + \lambda_0(\sum_{i=1}^{6} p_i - 1) + \lambda_1(\sum_{i=1}^{6} i \cdot p_i - 4.5) \tag{37}$$

After doing the classical optimization procedure, the desired probability would be:

$$p_i = exp(\lambda_0 + i\lambda_1) \tag{38}$$

$$\Lambda = exp(\lambda_0) = \sum_{i=1}^{6} exp(-i\lambda_1) \tag{39}$$

$$\Lambda = x(1-x)^{-1}(1-x^6) \quad \text{where:} \quad x = exp(-\lambda_1) \tag{40}$$

$$\frac{\partial \Lambda}{\partial \lambda_1} = -4.5 \quad \text{property of partition function} \tag{41}$$

$$3x^7 - 5x^6 + 9x - 7 = 0 \tag{42}$$

After obtaining the desired root for the above equation, the maximum entropy probabilities will be:

$$\{0.05435, 0.07877, 0.11416, 0.16545, 0.23977, 0.34749\} \tag{43}$$

While for iterative scaling process, the marginal construction function would be: $a_{1i} = \{1, 2, 3, 4, 5, 6\}$. According the iterative scaling process, the first estimation to the desired probability would be: p_i^0 : $\{1/6, 1/6, 1/6, 1/6, 1/6\}$. The first time of scaling would be:

$$p_1^1 = \mu\, p_1^0 (\frac{4.5}{3.5})^1$$
$$p_2^1 = \mu\, p_2^0 (\frac{4.5}{3.5})^2$$
$$p_3^1 = \mu\, p_3^0 (\frac{4.5}{3.5})^3$$
$$p_4^1 = \mu\, p_4^0 (\frac{4.5}{3.5})^4$$
$$p_5^1 = \mu\, p_5^0 (\frac{4.5}{3.5})^5$$
$$p_6^1 = \mu\, p_6^0 (\frac{4.5}{3.5})^6$$

After doing normalization, the estimation results from IS approach would be:

$$\{0.08123, 0.10444, 0.13428, 0.17265, 0.22198, 0.28540\}$$

All the ten times iteration results are listed in table 1 comparing the Lagrange result and the IS result, they are almost exactly the same. But the IS process is more straightforward and transparent.

Expert interpretation problem

It can still obtain the analytical equation in the dice problem; however, in general it is not possible. Suppose that an expert needs to estimate the rock type proportion for one outcrop. At first situation, he only knows there are five rock types $\{mud, sand, limestone, dolomite, anhydrite\}$ in this outcrop and without further sedimentary environment analysis informed. In this case, without more information in hand, the most intuitively appealing estimated proportion would be:

$$p(mud) = p_1 = 1/5$$
$$p(sand) = p_2 = 1/5$$
$$p(limestone) = p_3 = 1/5$$
$$p(dolomite) = p_4 = 1/5$$
$$p(anhydrite) = p_5 = 1/5$$

iteration time	p_1	p_2	p_3	p_4	p_5	p_6
1	0.08123	0.10444	0.13428	0.17265	0.22198	0.28540
2	0.06503	0.08945	0.12306	0.16928	0.23286	0.32032
3	0.05914	0.08365	0.11832	0.16736	0.23672	0.33482
4	0.05660	0.08108	0.11615	0.16639	0.23835	0.34143
5	0.05543	0.07989	0.11512	0.16591	0.23909	0.34456
6	0.05488	0.07931	0.11463	0.16567	0.23944	0.34606
7	0.05461	0.07904	0.11439	0.16556	0.23961	0.34680
8	0.05448	0.07890	0.11427	0.16550	0.23970	0.34715
9	0.05441	0.07883	0.11421	0.16547	0.23974	0.34733
10	0.05438	0.07880	0.11419	0.16546	0.23976	0.34741

Table 1: Ten times iteration results for the dir problem

Suppose after knowing the sedimentary background, it is known that either *mud* and *sand* would have a 30% proportion to exist. And also, in half the cases, the expert expects *mud* and *limestone* would be exist from the out crop. These two piece of information are incorporated into the estimation as two constraints:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/2 \end{bmatrix} \tag{44}$$

From the ME approach, the objective function would be:

$$L = \sum_{\ell=1}^{5} p_\ell \log p_\ell + \lambda_0 (\sum_{\ell=1}^{5} p_\ell - 1) + \sum_{m=1}^{2} \lambda_m (\sum_{\ell=1}^{5} a_{m\ell} p_\ell) \tag{45}$$

Using the classical derivative approach, the final expression for the desired probability would be:

$$p_\ell = exp(\lambda_0 + \sum_{m=1}^{2} \lambda_m a_{m\ell}) \tag{46}$$

Then the partition function would be written as:

$$\Lambda = exp(\lambda_0) = \sum_{\ell=1}^{5} exp(- \sum_{m=1}^{2} \lambda_m a_{m\ell}) \tag{47}$$

Theoretically, one set of nonlinear equation set can be obtained as:

$$\frac{\partial \lambda_0}{\partial \lambda_1} = \frac{\partial \log(\sum_{\ell=1}^{5} exp(- \sum_{m=1}^{2} \lambda_m a_{m\ell}))}{\partial \lambda_1} = -1/3 \tag{48}$$

$$\frac{\partial \lambda_0}{\partial \lambda_2} = \frac{\partial \log(\sum_{\ell=1}^{5} exp(- \sum_{m=1}^{2} \lambda_m a_{m\ell}))}{\partial \lambda_2} = -1/2 \tag{49}$$

Obtaining the solution for them is not easy.

While using IS, the probability estimation process would be more simple. The initial estimation

iteration time	p_1	p_2	p_3	p_4	p_5
1	0.2033	0.1626	0.2439	0.1951	0.1951
2	0.2034	0.1455	0.2679	0.1916	0.1916
3	0.2036	0.1373	0.2807	0.1892	0.1892
4	0.2039	0.1332	0.2875	0.1877	0.1877
5	0.2042	0.1311	0.2911	0.1868	0.1868
6	0.2044	0.1299	0.2931	0.1863	0.1863
7	0.2045	0.1294	0.2941	0.186	0.186
8	0.2046	0.129	0.2947	0.1859	0.1859
9	0.2046	0.1289	0.295	0.1858	0.1858
10	0.2046	0.1288	0.2951	0.1857	0.1857

Table 2: Ten times iteration results for the expert translator problem

would be $\{1/5, 1/5, 1/5, 1/5, 1/5\}$, the first iteration process is:

$$p_1^1 = \mu\, p_1^0 \,(\frac{1/3}{2/5})^1 (\frac{1/2}{2/5})^1$$
$$p_2^1 = \mu\, p_2^0 \,(\frac{1/3}{2/5})^1 (\frac{1/2}{2/5})^0$$
$$p_3^1 = \mu\, p_3^0 \,(\frac{1/3}{2/5})^0 (\frac{1/2}{2/5})^1$$
$$p_4^1 = \mu\, p_4^0 \,(\frac{1/3}{2/5})^0 (\frac{1/2}{2/5})^0$$
$$p_5^1 = \mu\, p_5^0 \,(\frac{1/3}{2/5})^0 (\frac{1/2}{2/5})^0$$

After ten times iteration, the constraints are very closely satisfied as shown in table 2.

4 Conclusion

The principle of Minimum Relative Entropy (MRE) is a more general form of the ME method. MRE has all the important attributes of the ME method with the advantage of more easy integration of the prior probability distribution. While used in discrete probability distribution estimation, there are some challenges. The iterative scaling solution to the MRE principle in discrete multivariate probability estimation is more straightforward and easy to implement. Using the iterative scaling makes its industrial implementation possible.

References

[1] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(6):620–630, 1957.

[2] E. T. Jaynes. Information theory and statistical mechanics. ii. *Physical Review*, 108(2):171–190, 1957.

[3] Edwin T. Jaynes. Where do we stand on maximum entropy? In Raphael D. Levine and Myron Tribus, editors, *The maximum entropy formalism*. The MIT Press, May 1978.

[4] H. Ku and S. Kullback. Approximating discrete probability distributions. *Information Theory, IEEE Transactions on*, 15(4):444–447, Jul 1969.

[5] S. Kullback. Letter to the editor: The kullback-leibler distance. *The American Statistician*, 41(4):340–341, 1987.

[6] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[7] S. Kullback and M.A. Khairat. A note on minimum discrimination information. *The Annals of Mathematical Statistics*, 37:279–280, 1966.

[8] Jone E. Shore and Rodney W. Johnson. Axiomatic derivation of the principle of maximum entropy and principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26:26–37, 1980.

[9] Harry H. Ku and Solomon Kullback. Loglinear models in contingency table analysis. *The American Statistician*, 28(4):115–122, 1974.

[10] Yvonne M. M. Bishop. Full contingency tables, logits, and split contingency tables. *Biometrics*, 25(2):383–399, 1969.

[11] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[12] D. V. Gokhale. Analysis of log-linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(3):371–376, 1972.

[13] I. J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934, 1963.

[14] P.M. Lewis. Approximating probability distributions to reduce storage requirement. *Information and control*, 2:214 to 225, 1959.