axioms

MDPI

Article

Parameter Estimation of the Dirichlet Distribution Based on Entropy

Büşra Şahin [1,†], Atıf Ahmet Evren [2,†], Elif Tuna [2,†], Zehra Zeynep Şahinbaşoğlu [2,*,†] and Erhan Ustaoğlu [3,†]

[1] Department of Computer Engineering, Faculty of Engineering, Halic University, Eyupsultan, 34060 Istanbul, Turkey; busrasahin@halic.edu.tr
[2] Department of Statistics, Faculty of Sciences and Literature, Yildiz Technical University, Davutpasa, Esenler, 34210 Istanbul, Turkey; aevren@yildiz.edu.tr (A.E.); eozturk@yildiz.edu.tr (E.T.)
[3] Department of Informatics, Faculty of Management, Marmara University, Göztepe, 34180 Istanbul, Turkey; erhan.ustaoglu@marmara.edu.tr
* Correspondence: zeynepsahinbasoglu@gmail.com
† These authors contributed equally to this work.

Abstract: The Dirichlet distribution as a multivariate generalization of the beta distribution is especially important for modeling categorical distributions. Hence, its applications vary within a wide range from modeling cell probabilities of contingency tables to modeling income inequalities. Thus, it is commonly used as the conjugate prior of the multinomial distribution in Bayesian statistics. In this study, the parameters of a bivariate Dirichlet distribution are estimated by entropy formalism. As an alternative to maximum likelihood and the method of moments, two methods based on the principle of maximum entropy are used, namely the ordinary entropy method and the parameter space expansion method. It is shown that in estimating the parameters of the bivariate Dirichlet distribution, the ordinary entropy method and the parameter space expansion method give the same results as the method of maximum likelihood. Thus, we emphasize that these two methods can be used alternatively in modeling bivariate and multinomial Dirichlet distributions.

Keywords: Dirichlet distribution; principle of maximum entropy; ordinary entropy method; parameter space expansion method; method of moments; maximum likelihood estimation

MSC: 62H12; 94A17; 1F66; 54C70

check for updates

Citation: Şahin, B.; Evren, A.A.; Tuna, E.; Şahinbaşoğlu, Z.Z.; Ustaoğlu, E. Parameter Estimation of the Dirichlet Distribution Based on Entropy. *Axioms* **2023**, *12*, 947. https://doi.org/10.3390/axioms12100947

Academic Editor: Lechang Yang, Qingqing Zhai, Rui Peng and Aibo Zhang

Received: 31 July 2023
Revised: 28 September 2023
Accepted: 30 September 2023
Published: 5 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In statistics, the method of moments and maximum likelihood are used frequently, details of which can be found in [1,2]. For a long time, their asymptotic properties have been studied in detail [3]. Since the asymptotic distributions of estimators found by these two methods are normal, they have been proven to be very powerful tools for parameter estimation. However, nowadays, alternative estimation methods based on entropy maximization are applied increasingly frequently.

In 1948, [4] defined entropy as a numerical measure of uncertainty, or conversely the information content, associated with a probability distribution $f(x;\theta)$ with parameter θ. It is used to describe a random variable X and is mathematically expressed as

$$I[f] = -\int_{-\infty}^{\infty} f(x;\theta) ln f(x;\theta) dx, \ \int_{-\infty}^{\infty} f(x;\theta) dx = 1 \quad (1)$$

for continuous X, where $I[f]$ can be considered the mean value of $lnf(x;\theta)$. For discrete probability distributions, the integration operator in (1) is simply replaced by the sum-

mation operator. Rényi (1961) provided a generalization of Shannon entropy [5]. Rényi entropy is also called α-class entropy. For a discrete case, it is defined as

$$H_R = \frac{ln(\sum_{i=1}^{K} p_i^{\alpha})}{1-\alpha} \; for \; \alpha > 0 \; and \; \alpha \neq 1 \tag{2}$$

By L'Hôspital's rule

$$\lim_{\alpha \to 1} \frac{\frac{d}{d\alpha}(ln(\sum_{i=1}^{K} p_i^{\alpha}))}{\frac{d}{d\alpha}(1-\alpha)} = \lim_{\alpha \to 1} \frac{\frac{\sum_{i=1}^{K} p_i^{\alpha} lnp_i}{\sum_{i=1}^{K} p_i^{\alpha}}}{-1} = -\sum_{i=1}^{K} p_i lnp_i = H_S \tag{3}$$

Therefore, Shannon entropy can be evaluated as a special case of Rényi entropy. Another generalization of Shannon entropy was realized by Constantino Tsallis (1988) [5]. Tsallis entropy is also known as β-class entropy [6]. It is defined as

$$H_T = \frac{1-\sum_{i=1}^{K} p_i^{\alpha}}{\alpha - 1} \; for \; \alpha > 0 \; and \; \alpha \neq 1 \tag{4}$$

By L'Hôspital's rule,

$$\lim_{\alpha \to 1} \frac{1-\sum_{i=1}^{K} p_i^{\alpha}}{\alpha - 1} = \lim_{\alpha \to 1} \frac{\frac{d}{d\alpha}(1-\sum_{i=1}^{K} p_i^{\alpha})}{\frac{d}{d\alpha}(\alpha - 1)} = \lim_{\alpha \to 1} \frac{-\sum_{i=1}^{K} p_i^{\alpha} lnp_i}{1} = -\sum_{i=1}^{K} p_i lnp_i \tag{5}$$

In other words, Tsallis entropy approaches Shannon entropy as $\alpha \to 1$ as well as Rényi entropy. Note that for continuous distributions, the summation signs in defining equations are replaced by integration signs.

Kullback (1959) used entropy and relative entropy as the two key concepts in multivariate statistical analysis [7]. Asymptotic distributions of various entropy measures can be found in [8]. Pardo emphasizes that entropy and relative entropy formulas can be derived as special cases of divergence measures [9].

Entropy-Based Parameter Estimation in Hydrology is the first book to focus on parameter estimation using entropy for a number of distributions frequently used in hydrology [10], including uniform, exponential, normal, two-parameter lognormal, extreme value type I, Weibull, gamma, Pearson, and two-parameter Pareto distributions, among others. Singh also applies entropy theory to some problems of hydraulic and environmental engineering [11–13].

The principle of maximum entropy (POME), described by Jaynes as "the least biased estimate possible on the given information", can be stated mathematically as follows [14]: Given m linearly independent constraints C_i in the form

$$C_i = \int_a^b y_i(x) f(x) dx, \; i = 1, 2, \ldots, m, \tag{6}$$

where $y_i(x)$ are some functions whose averages over $f(x)$ are specified, the maximum of I, subject to the conditions in Equation (6), is given by the distribution

$$f(x) = \exp\big[-\lambda_0 - \sum_{i=1}^{m} \lambda_i y_i(x)\big], \tag{7}$$

where $\lambda_i, i = 0, 1, \ldots, m$ are Lagrange multipliers and can be determined from Equations (6) and (7) along with the normalization condition in Equation (1).

The general procedure for entropy-based parameter estimation involves (1) defining given information in terms of constraints, (2) maximizing entropy subject to given information, and (3) relating parameters to the given information. In this procedure, Lagrange multipliers are related to the constraints on one hand and to the distribution parameters on

the other. One can eliminate the Lagrange multipliers and obtain parameter estimations as well.

The parameter space expansion method was developed by Singh and Rajogopal (1986). This method is different from the previous entropy method in that it employs enlarged parameter space and maximizes entropy subject to both the parameters and the Lagrange multipliers [15]. The method works as follows: for the given distribution, first the constraints are defined, and the POME formulation is obtained in terms of the parameters to be estimated and the Lagrange multipliers. After the maximization procedure, the parameter estimations can be obtained.

Entropy-based models have been intensively used for determining parameter estimations in recent years. For example, Song and Kang examined two entropy-based methods that both use the POME for the estimation of the parameters of the four-parameter exponential gamma distribution [16]. Hao and Singh applied two entropy-based methods, also using the POME, for the estimation of the parameters of the extended Burr XII distribution [17]. Singh and Deng revisited the four-parameter kappa distribution, presented an entropy-based method for estimating its parameters, and compared its performance with that of maximum likelihood estimation, methods of moments, and L moments [18]. Gao and Han used the maximum entropy method to apply a concrete solution to a special nonlinear expectation problem in a special parameter space and analyzed the convergence for the maximum entropy solution [19].

The objective of the present paper is to apply ordinary entropy and parameter space expansion to estimate the parameters of a bivariate Dirichlet distribution as an alternative to the known methods, and then to compare them with those estimated by the maximum likelihood method and method of moments.

2. Dirichlet Distribution

The beta distribution plays an important role in Bayesian statistics, especially in modeling the parameters of the Bernoulli distribution [20]. The Dirichlet distribution is a multivariate generalization of the beta distribution. Thus, the Dirichlet distribution and the generalized Dirichlet distribution can both be used as a conjugate prior for a multinomial distribution [21].

Let $X^k = [X_1, X_2, \ldots, X_k]$ be a vector with k components, $X_i \geq 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} x_i = 1$. Also, $a^k = [a_1, a_2, \ldots, a_k]$, where $a_i > 0$ for each i. The probability density function (pdf) of the Dirichlet distribution is given as

$$f(x^k) = \frac{\Gamma(a_0)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} x_i^{a_i - 1}, \tag{8}$$

where $a_0 = \sum_{i=1}^{k} a_i$, $x_i > 0$, $x_1 + x_2 + \cdots + x_{k-1} < 1$, and $x_k = 1 - x_1 - \cdots - x_{k-1}$ and Γ is the Euler's gamma function, which is denoted by the formula $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ or $\Gamma(x) = (x-1)!$.

It can be noted that marginals of this Dirichlet distribution are beta distributions [22], namely $X_i \sim Beta\left(a_i, (\sum_{j=1}^{k} a_j) - a_i\right)$. The moments are given by

$$E[X_i] = \frac{a_i}{a_0} \tag{9}$$

$$Var[X_i] = \frac{a_i(a_0 - a_i)}{a_0^2(a_0 + 1)} \tag{10}$$

$$Cov(X_i, X_j) = -\frac{a_i a_j}{a_0^2(a_0 + 1)} \tag{11}$$

$$Cor(X_i, X_j) = \sqrt{\frac{a_i a_j}{(a_0 - a_i)(a_0 - a_j)}} \tag{12}$$

For further properties, one may refer to [22–24].

3. Ordinary Entropy Method

In the ordinary entropy method, there are three steps in parameter estimation: (1) specification of appropriate constraints, (2) derivation of the entropy function of the distribution, and (3) derivation of the relations between parameters and constraints.

3.1. Specification of Constraints

Taking the natural logarithm of Equation (8), we obtain

$$lnf(x^k) = ln\Gamma(a_0) - ln\left(\prod_{i=1}^{k}\Gamma(a_i)\right) + \sum_{i=1}^{k} ln(x_i^{a_i-1}) \tag{13}$$

Multiplying Equation (13) by $[-f(x^k)]$ and integrating between $[0,1]$ and $[0, 1-x_i]$, we obtain the entropy function

$$I[f] = -\int\cdots\int f(x^k)lnf(x^k)dx_1\ldots dx_{k-1} = \left[-ln\frac{\Gamma(a_0)}{\prod_{i=1}^{k}\Gamma(a_i)}\right]\int\cdots\int f(x^k)dx_1\ldots dx_{k-1}$$

$$-\int\cdots\int\sum_{i=1}^{k} ln(x_i^{a_i-1})f(x^k)dx_1\ldots dx_{k-1} \tag{14}$$

To maximize $I[f]$ in Equation (14), the following constraints should be satisfied:

$$\int\cdots\int f(x^k)dx_1\ldots dx_{k-1} = 1 \tag{15}$$

$$\int\cdots\int lnx_i f(x^k)dx_1\ldots dx_{k-1} = E[lnx_i],\ i = 1,\ldots,k-1 \tag{16}$$

$$\int\cdots\int ln(1 - x_1 - \cdots - x_{k-1})f(x^k)dx_1\ldots dx_{k-1} = E[1 - x_1 - \cdots - x_{k-1}] \tag{17}$$

3.2. Construction of the Partition Function and Zeroth Lagrange Multiplier

The least biased pdf, $f(x^k)$ consistent with equations from (15) to (17) and by POME, takes the following form:

$$f(x^k) = exp[-\lambda_0 - \sum_{i=1}^{k-1}\lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})], \tag{18}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_k$ are Lagrange multipliers. Substituting (18) in (15) yields

$$\int\cdots\int \exp[-\lambda_0 - \sum_{i=1}^{k-1}\lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})]dx_1\ldots dx_{k-1} = 1 \tag{19}$$

Equation (19) gives the partition function as

$$exp(\lambda_0) = \int\cdots\int \exp[-\sum_{i=1}^{k-1}\lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})]dx_1\ldots dx_{k-1}, \tag{20}$$

which may be further simplified as follows:

$$exp(\lambda_0) = \int \cdots \int \prod_{i=1}^{k-1} x_i^{-\lambda_i} (1 - x_1 - \cdots - x_{k-1})^{-\lambda_k} dx_1 \ldots dx_k = \frac{\prod_{i=1}^{k} \Gamma(1-\lambda_i)}{\Gamma(k - \lambda_1 - \ldots \lambda_k)} \quad (21)$$

The zeroth Lagrange multiplier λ_0 is obtained from Equation (21) as

$$\lambda_0 = \sum_{i=1}^{k} ln\Gamma(1-\lambda_i) - ln\Gamma(k - \lambda_1 - \ldots \lambda_k) \quad (22)$$

The zeroth Lagrange multiplier is also obtained from (20) as

$$\lambda_0 = ln \int \cdots \int \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1} \quad (23)$$

3.3. Relation between Lagrange Multipliers and Constraints

Differentiating Equation (23) with respect to $\lambda_1, \ldots, \lambda_k$, we obtain the derivatives of λ_0 with respect to $\lambda_1, \ldots, \lambda_k$:

$$\frac{\partial \lambda_0}{\partial \lambda_1} = -\frac{\int \cdots \int \ lnx_1 \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}}{\int \cdots \int \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}}$$

$$= -\int \cdots \int \ lnx_1 \exp[-\lambda_0 - \sum_{i=1}^{k-1} -\lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}$$

$$= -E[lnX_1] \quad (24)$$

$$\frac{\partial \lambda_0}{\partial \lambda_2} = -\frac{\int \cdots \int \ lnx_2 \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}}{\int \cdots \int \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}}$$

$$= -\int \cdots \int \ lnx_2 \exp[-\lambda_0 - \sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}$$

$$= -E[lnX_2] \quad (25)$$

If it continues like this until $k-1$,

$$\frac{\partial \lambda_0}{\partial \lambda_{k-1}} = -\frac{\int \cdots \int \ lnx_{k-1} \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}}{\int \cdots \int \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}}$$

$$= -\int \cdots \int \ lnx_{k-1} \exp[-\lambda_0 - \sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})] dx_1 \ldots dx_{k-1}$$

$$= -E[lnX_{k-1}] \quad (26)$$

Furthermore,

$$\frac{\partial \lambda_0}{\partial \lambda_k} = -\frac{\int \cdots \int \ ln(1 - (\sum_{i=1}^{k-1} x_i)) \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - (\sum_{i=1}^{k-1} x_i))] dx_1 \ldots dx_{k-1}}{\int \cdots \int \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - (\sum_{i=1}^{k-1} x_i))] dx_1 \ldots dx_{k-1}}$$

$$= \int \cdots \int \ ln(1 - x_1 - \cdots - x_{k-1}) \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1 - (\sum_{i=1}^{k-1} x_i))] dx_1 \ldots dx_{k-1}$$

$$= -E[ln(1 - x_1 - \cdots - x_{k-1})] \quad (27)$$

Differentiating Equation (22) with respect to $\lambda_1, \lambda_2, \ldots, \lambda_k$, we obtain

$$\frac{\partial \lambda_0}{\partial \lambda_1} = \frac{\partial(\sum_{i=1}^{k} ln\Gamma(1-\lambda_i) - ln\Gamma(k-\lambda_1-\ldots\lambda_k))}{\partial \lambda_1} = -\psi(1-\lambda_1) + \psi(k-\lambda_1-\ldots\lambda_k) \tag{28}$$

$$\frac{\partial \lambda_0}{\partial \lambda_2} = \frac{\partial(\sum_{i=1}^{k} ln\Gamma(1-\lambda_i) - ln\Gamma(k-\lambda_1-\ldots\lambda_2))}{\partial \lambda_2} = -\psi(1-\lambda_2) + \psi(k-\lambda_1-\ldots\lambda_k) \tag{29}$$

Similarly, for the k term,

$$\frac{\partial \lambda_0}{\partial \lambda_k} = \frac{\partial(\sum_{i=1}^{k} ln\Gamma(1-\lambda_i) - ln\Gamma(k-\lambda_1-\ldots\lambda_k))}{\partial \lambda_k} = -\psi(1-\lambda_k) + \psi(k-\lambda_1-\ldots\lambda_k), \tag{30}$$

where $\psi(x)$ is the digamma function, which is defined as $\psi(x) = \frac{d}{dx} ln(\Gamma(x))$ [25].

By equating (24) and (28), we obtain

$$E[lnX_1] = \psi(1-\lambda_1) - \psi(k-\lambda_1-\ldots\lambda_k) \tag{31}$$

Secondly , by equating (25) and (29), we obtain

$$E[lnX_2] = \psi(1-\lambda_2) - \psi(k-\lambda_1-\ldots\lambda_k) \tag{32}$$

If we go on until the $(k-1)$ term

$$E[lnX_{k-1}] = \psi(1-\lambda_{k-1}) - \psi(k-\lambda_1-\ldots\lambda_k) \tag{33}$$

Next, by (27) and (30), we obtain

$$E[ln(1-x_1-\cdots-x_{k-1})] = \psi(1-\lambda_k) - \psi(k-\lambda_1-\ldots\lambda_k) \tag{34}$$

3.4. Relation between Lagrange Multipliers and Parameters

Substituting (22) into (18) yields

$$f(x^k) = exp[-\sum_{i=1}^{k} ln\Gamma(1-\lambda_i) + ln\Gamma(k-\lambda_1-\ldots\lambda_k) - \sum_{i=1}^{k-1} \lambda_i lnx_i - \lambda_k ln(1-(\sum_{i=1}^{k-1} x_i))] \tag{35}$$

$$= \frac{\Gamma(k-\lambda_1-\ldots\lambda_k)}{\prod_{i=1}^{k}\Gamma(1-\lambda_i)} \prod_{i=1}^{k} x_i^{-\lambda_i}$$

A comparison of Equation (35) with Equation (8) shows that

$$a_i = 1-\lambda_i \tag{36}$$

3.5. Relation between Parameters and Constraints

The parameters of the Dirichlet distribution are related to the Lagrange multipliers. In turn, these parameters are related to the known constraints by equations from (31) to (34). By eliminating Lagrange multipliers from these sets of equations, we can obtain an alternative way of presentation as shown below:

$$E[lnX_1] = \psi(a_1) - \psi(a_0) \tag{37}$$

$$E[lnX_2] = \psi(a_2) - \psi(a_0) \tag{38}$$

$$\vdots$$

$$E[lnX_{k-1}] = \psi(a_{k-1}) - \psi(a_0) \tag{39}$$

$$E[ln(1 - x_1 - \cdots - x_{k-1})] = \psi(a_k) - \psi(a_0) \tag{40}$$

3.6. Distribution Entropy

From (14),

$$I[f] = -\int \cdots \int f(x^k) lnf(x^k) dx_1 \ldots dx_{k-1}$$

$$= \left[-ln\frac{\Gamma(a_0)}{\prod_{i=1}^{k}\Gamma(a_i)} \right] - \int \cdots \int \sum_{i=1}^{k} ln(x_i^{a_i-1}) f(x^k) dx_1 \ldots dx_{k-1} \tag{41}$$

$$= \left[ln\frac{\prod_{i=1}^{k}\Gamma(a_i)}{\Gamma(a_0)} \right] - \sum_{i=1}^{k}(a_i - 1) lnx_i$$

4. Parameter Space Expansion Method

4.1. Specification of Constraints

Following [15], the constraints for this method are Equation (15) and

$$\int \cdots \int lnx_i^{a_i-1} f(x^k) dx_1 \ldots dx_{k-1} = E[lnX_i^{a_i-1}],\ i = 1, \ldots, k-1 \tag{42}$$

$$\int \cdots \int ln(1 - x_1 - \cdots - x_{k-1})^{a_k-1} f(x^k) dx_1 \ldots dx_{k-1} = E[(1 - x_1 - \cdots - x_{k-1})^{a_k-1}] \tag{43}$$

4.2. Derivation of the Entropy Function

The pdf that corresponds to the POME and that is consistent with Equation (15), (42), and (43) takes the form

$$f(x^k) = exp[-\lambda_0 - \sum_{i=1}^{k-1} \lambda_i lnx_i^{a_i-1} - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})^{a_k-1}], \tag{44}$$

where $\lambda_0, \lambda_1, \ldots, \lambda_k$ are Lagrange multipliers. Substituting (44) into Equation (15) yields

$$exp(\lambda_0) = \int \cdots \int \exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i^{a_i-1} - \lambda_k ln(1 - x_1 - \cdots - x_{k-1})^{a_k-1} dx_1 \ldots dx_{k-1}]$$

$$= \frac{\prod_{i=1}^{k}\Gamma(1 - \lambda_i(a_i - 1))}{\Gamma(k - \lambda_1(a_1 - 1) - \ldots \lambda_k(a_k - 1))} \tag{45}$$

Substitution of Equation (45) into (44) gives

$$f(x^k) = \frac{\Gamma(k - \lambda_1(a_1 - 1) - \ldots \lambda_k(a_k - 1))}{\prod_{i=1}^{k}\Gamma(1 - \lambda_i(a_i - 1))} exp[-\sum_{i=1}^{k-1} \lambda_i lnx_i^{a_i-1} - \lambda_k ln(1 - (\sum_{i=1}^{k-1} x_i))^{a_k-1}] \tag{46}$$

A comparison of Equation (46) with Equation (15) shows that $\lambda_1 = \ldots \lambda_k = -1$ and taking the logarithm of (46) and multiplying by $[-f(x^k)]$and integrating between $[0, 1]$ and $[0, 1 - x_i]$, we obtain the entropy function

$$I[f] = -ln\Gamma(k - \lambda_1(a_1 - 1) - \ldots \lambda_k(a_k - 1)) + \sum_{i=1}^{k}\Gamma(1 - \lambda_i(a_i - 1)) + \sum_{i=1}^{k} \lambda_i E[lnx_i^{a_i-1}] \tag{47}$$

4.3. Relation between Parameters and Constraints

Equating the partial derivatives of (47) with respect to $\lambda_1, \ldots, \lambda_k, a_1, \ldots, a_k$ to zero, one obtains

$$\frac{\partial\, I[f]}{\partial \lambda_1} = 0 = (a_1 - 1)\psi(K_1) - (a_1 - 1)\psi(K_2) + E[lnX_1^{(a_1-1)}] \tag{48}$$

$$K_1 = (k - \lambda_1(a_1 - 1) - \ldots \lambda_k(a_k - 1)),\ K_2 = (1 - \lambda_1(a_1 - 1))$$

$$\vdots$$

$$\frac{\partial\, I[f]}{\partial \lambda_k} = 0 = (a_k - 1)\psi(K_1) - (a_k - 1)\psi(K_k) + E[lnX_k^{(a_k-1)}] \tag{49}$$

$$K_k = (k - \lambda_k(a_k - 1))$$

$$\frac{\partial\, I[f]}{\partial a_1} = 0 = \lambda_1\psi(K_1) - \lambda_1\psi(K_2) + \lambda_1 E[lnX_1] \tag{50}$$

$$\vdots$$

$$\frac{\partial\, I[f]}{\partial a_k} = 0 = \lambda_k\psi(K_1) - \lambda_k\psi(K_k) + \lambda_k E[lnX_k] \tag{51}$$

The simplification of equations from (48) to (51) yields

$$E[lnX_1] = \psi(K_2) - \psi(K_1) \tag{52}$$

$$\vdots$$

$$E[lnX_k] = \psi(K_k) - \psi(K_1) \tag{53}$$

These equations provide the parameter estimators of the Dirichlet distribution.

5. Two Other Parameter Estimation Methods

5.1. Method of Moments

The Dirichlet distribution has k parameters like a_i , $i = 1, \ldots, k$. Therefore, i moments are needed for the parameter estimation. We have the moments, variances, and covariance formula in (9), (10), and (11). Because of (11),

$$E(X_iX_j) = Cov(X_i, X_j) + E(X_i)E(X_j) = \frac{a_i a_j}{a_o(a_0 + 1)} \tag{54}$$

$$\frac{E(X_iX_j)}{Cov(X_i, X_j)} = -a_0 \tag{55}$$

If we multiply (55) by the negative of (9),

$$\hat{a_i} = -E(X_i)\frac{E(X_iX_j)}{Cov(X_i, X_j)} \tag{56}$$

Due to (55) and (56), the last parameter estimation is

$$\hat{a_k} = \frac{-E(X_iX_j)(1 - E(X_1) - \ldots E(X_{k-1})}{Cov(X_i, X_j)} \tag{57}$$

5.2. *Method of Maximum Likelihood Estimation*

The likelihood function L, where n is the sample size, is

$$L = \left[\frac{\Gamma(a_0)}{\prod_{i=1}^{k}\Gamma(a_i)}\right]^n \prod_{j=1}^{n}\left(\prod_{i=1}^{k} x_{ij}^{a_i-1}\right) \tag{58}$$

Then the log likelihood function, lnL, is

$$lnL = nln\Gamma(a_0) - n\sum_{i=1}^{k} ln\Gamma(a_i) + \sum_{j=1}^{n} ln\left[\prod_{i=1}^{k} x_{ij}^{a_i-1}\right] \tag{59}$$

Differentiating Equation (59) with respect to parameters $a_1, \ldots, a_k$, respectively, and equating each derivative to zero yield the following equations:

$$E[lnX_1] = \psi(a_1) - \psi(a_0) \tag{60}$$

$$E[lnX_2] = \psi(a_2) - \psi(a_0) \tag{61}$$

$$\vdots$$

$$E[lnX_{k-1}] = \psi(a_{k-1}) - \psi(a_0) \tag{62}$$

$$E[ln(1 - x_1 - \cdots - x_{k-1})] = \psi(a_k) - \psi(a_0) \tag{63}$$

These results are the same as those found by the ordinary entropy method and parameter space expansion method.

The maximum likelihood (ML) estimation method provides singular point estimates for model parameters while overlooking the residual uncertainty inherent in the estimation process. Conversely, the Bayesian estimation method adopts a different approach, yielding posterior probability distributions encompassing the entire spectrum of model parameters. This is achieved by integrating the observed data with prior distributions [26]. Broadly speaking, when contrasted with ML estimation, Bayesian parameter estimation within a statistical model has the potential to yield a robust and stable estimate. This is primarily due to its ability to incorporate the accompanying uncertainty into the estimation process, a particularly valuable attribute when dealing with limited amounts of observed data [27]. The Dirichlet distribution, being a constituent of the exponential family, possesses a corresponding conjugate prior. Nevertheless, due to the intricate nature of the posterior distribution, its practical utility in problem-solving scenarios is limited. Consequently, the task of Bayesian estimation for the Dirichlet distribution, in a general context, lacks analytical tractability. To achieve this objective, Zao employed an approximation approach to model the parameter distribution within the Dirichlet distribution. Specifically, they approximated it with a multivariate Gaussian distribution, leveraging the expectation propagation (EP) framework [28]. Furthermore, there are some studies in reliability engineering which estimate parameters with the determination of quantiles by the application of the maximum likelihood method, such as [29].

6. Simulation and Comparison of Parameter Estimation Methods

Simulation from the Dirichlet distribution can be performed in two steps: the probability integral theorem states that the distribution function of any continuous distribution is uniform on $(0, 1)$. Then, by the inverse distribution function of gamma, one may simulate a number of independent gamma variates as needed. In other words, first one may simulate k independent gamma variates $X_1, X_2, \ldots, X_n$ such that $X_i \sim Gamma(\alpha_i, 1), \quad i = 1, 2, \ldots, k$

and calculate $Y_j = \frac{X_j}{\sum_{i=1}^{k} X_i}; j = 1, 2, \ldots, k$. Then the random vector $(Y_1, Y_2 \ldots, Y_k)$ fits the Dirichlet distribution, having parameter vector $(\alpha_1, \alpha_2, \ldots, \alpha_k)$ [30]. This procedure can easily be realized even by Microsoft Excel.

In the present study, we first simulated 1000 (X, Y) pairs from the Dirichlet distribution for some arbitrary parameters α_1, α_2 and α_3. We obtained estimates obtained by four methods but since the maximum likelihood estimators and estimators obtained by the ordinary entropy method and parameter space expansion method are all the same, the comparison is between moment estimates and the rest. Then we repeated this experiment 5000 times. The summarizing statistics are as shown in Table 1:

Table 1. Results of some simulations (1000 runs, 5000 runs).

1000 runs/5000 runs	$\alpha_1 = 3$	$\alpha_2 = 2$	$\alpha_3 = 4$
MOM	2.79/2.92	1.86/1.94	3.81/3.88
APE	6.75/2.35	6.99/2.9	4.73/2.94
MLE	3.06/3.23	2.21/2.11	4.64/4.32
APE	2.07/7.85	10.5/5.98	16.24/8.1
1000 runs/5000 runs	$\alpha_1 = 4$	$\alpha_2 = 0.25$	$\alpha_3 = 2$
MOM	3.96/3.94	0.25/0.24	1.96/1.97
APE	0.99/1.26	2.82/0.05	1.53/1.18
MLE	4.28/4.02	0.26/0.22	2.03/1.96
APE	7.01/0.57	4.27/9.69	1.85/1.79
1000 runs/5000 runs	$\alpha_1 = 0.5$	$\alpha_2 = 3$	$\alpha_3 = 2$
MOM	0.48/0.5	3.1/3.09	2.09/2.04
APE	3.91/1.94	3.36/3.13	4.67/2.09
MLE	0.61/0.5	3.39/3.1 1	2.32/2.06
APE	22.48/0.58	13.01/3.92	16.1/3.34
1000 runs/5000 runs	$\alpha_1 = 3$	$\alpha_2 = 3$	$\alpha_3 = 4$
MOM	3.01/3.04	3.05/2.99	4.13/4.04
APE	0.49/1.63	1.82/0.36	3.38/1.04
MLE	3.22/2.86	3.41/2.88	4.61/3.61
APE	7.57/4.38	13.67/3.71	15.48/9.64
1000 runs/5000 runs	$\alpha_1 = 13$	$\alpha_2 = 2$	$\alpha_3 = 0.75$
MOM	14.04/12.99	2.1/1.98	0.83/0.73
APE	8/0.02	5.34/0.56	11/1.45
MLE	12.52/13.11	1.96/2.02	0.75/0.69
APE	3.64/0.87	1.89/1.34	0.72/7.85

Note that the maximum likelihood estimates (MLEs) are obtained by Excel Solver. In general, moment estimates and MLEs are close to each other. Absolute percentage errors (APEs) are calculated by the following formula:

$$APE = \frac{100 * |parameter - estimate|}{parameter} \tag{64}$$

Then, it can be inferred that one measure does not dominate all the time, i.e., there are some instances in which moment estimators perform better than the others, and there are other instances in which maximum likelihood estimators do it better. In any case, it is definite from the table that increasing the number of simulations increases precision considerably.

Note that, by the central limit theorem, moment estimators are expected to be distributed normally for a large number of observations (or simulations) since a moment estimator considers a sum of random observations (or the sum of some power of these random observations). Maximum likelihood estimators also have the asymptotic normality

property with lower variances. For the Dirichlet distribution, we found that the entropy estimators mentioned above and maximum likelihood estimators are identical. Therefore, entropy estimators also have the asymptotic normality property.

Finally, we note that the selection of $(\alpha_1, \alpha_2, \alpha_3)$ is quite arbitrary just for addressing the fact that maximum likelihood estimators (and maximum entropy estimators) are better (i.e., show lower sampling variability as compared to moment estimators). Actually, this was not the case all the time in the simulations. Since, in our study, the initial estimates of maximum likelihood (and maximum entropy) are provided by the method of moments and since the initial estimates provided by moments are close enough to the actual parameters, a great improvement in sampling variability may not be achieved. This is probably due to the nature of nonlinear estimation. To have a better picture, first simulating a random vector $(\alpha_1, \alpha_2, \alpha_3)$ several times, then trying to calculate moment estimates and then, based on these initial estimates, moving forward to maximum likelihood estimates may be meaningful.

7. Conclusions

In the present study, parameter estimations of the Dirichlet distribution are obtained by four methods. For a Dirichlet distribution with three parameters, parameter estimates found by entropy methods (that we considered here) and by maximum likelihood are almost the same. Maximum likelihood estimators are consistent, most efficient, sufficient, tend to normality (as the sample size increases), and are invariant under functional transformations [31]. Therefore, parameter estimators found by the entropy methodology have the same appealing properties as the maximum likelihood estimators. Based on the fact that the sample moment will tend to be more concentrated for the corresponding population moment for larger samples, a sample moment can also be used to estimate population moments [1]. In general, moment estimators are asymptotically normally distributed and consistent. However, their variance may be larger than that of estimators derived by other methods [32]. Yet, it may be a good idea to start nonlinear estimation either for maximum likelihood or entropy maximization methods with initial moment estimates. In the present study, we started with moment estimates of a Dirichlet distribution with arbitrarily selected parameters to demonstrate that better parameter estimates (i.e., estimates both with lower bias and lower sampling variability) can be achieved. The simulation part of this work can be enlarged by determining parameters randomly by further simulations for further generalizations.

Author Contributions: All authors made the same contributions. All authors made the same contributions. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All results presented in the article were produced from model simulations. Therefore, there are no data to be made available. Researchers who wish to replicate the study should use Microsoft Excel and the parameters described in the article. With those parameters, researchers can use modeling simulations to replicate the tables and figures presented in the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mood, A.M.; Graybill, F.A.; Boes, D. *Introduction to the Theory of Statistics*; McGraw-Hill Edition: New York, NY, USA, 1974.
2. Casella, G.; Berger, R.L. *Statistical Inference*, 2nd ed.; Duxbury Advanced Series; Cengage Learning: Pacific Grove, CA, Australia, 2002.
3. Dasgupta, A. *Asymtotic Theory of Statistics and Probability*; Springer: New York, NY, USA, 2002.
4. Shannon, C.E. A mathematical theory of communication. *Bell. Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
5. Renyi, A. On measures of entropy and information. In Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960; Volume 1, pp. 547–561.
6. Ullah, A. Entropy, divergence and distance measures with econometric applications. *J. Stat. Plan. Inference* **1996**, *49*, 137–162. [CrossRef]
7. Kullback, S. *Information Theory and Statistics*; Dover Publications: New York, NY, USA, 1978.

8. Esteban, M.D.; Morales, D. A summary on entropy statistics. *Kybernetika* **1995**, *1*, 337–346.
9. Pardo, L. *Statistical İnference Based on Divergence Measures*; Chapman&Hall/CRC: New york, NY, USA, 2006.
10. Singh, V.P. *Entropy-Based Parameter Estimation in Hydrology*; Kluwer Academic Publishers: Boston, MA, USA, 1998.
11. Singh, V.P. *Entropy Theory and Its Application in Environmental and Water Engineering*; John Wiley and Sons: West Sussex, UK, 2013.
12. Singh, V.P. *Entropy Theory in Hydraulic Engineering: An Introduction*; ASCE Press: Reston, VA, USA, 2015.
13. Singh, V.P. *Entropy Theory in Hydrologic Science and Engineering*; Hill Education: New York, NY, USA, 2014.
14. Jaynes, E.T. *Probability Theory: The Logic of Science*; Cambridge University Press: Cambridge, UK, 2002.
15. Singh, V.P.; Rajagopal, A.K. A new method of parameter estimation for hydrologic frequency analysis. *Hydrol. Sci. Technol.* **1986**, *3*, 33–40.
16. Song, S.; Song, X.; Kang, Y. Entropy-Based Parameter Estimation for the Four-Parameter Exponential Gamma Distribution. *Entropy* **2017**, *19*, 189. [CrossRef]
17. Hao, Z.; Singh, V.P. Entropy-based parameter estimation for extended Burr XII distribution. *Stoch Environ. Res Risk Assess.* **2009**, *23*, 1113–1122. [CrossRef]
18. Singh, V.P.; Deng, Z.Q. Entropy-based parameter estimation for kappa distribution. *J. Hydrol. Eng.* **2003**, *8*, 81–92. [CrossRef]
19. Gao, L.; Han, D. Methods of Moment and Maximum Entropy for Solving Nonlinear Expectation. *Mathematics* **2019**, *7*, 45. [CrossRef]
20. De Groot, M.; Shervish, M. *Probability and Statistics*, 4th ed.; Addison-Wesley: Boston, MA, USA, 2002.
21. Press, S.J. *Applied Multivariable Analysis Using Bayesian and Frequent Methods of Inference*; Dover Publications: Mineola, NY, USA, 1981.
22. Lin, J. On the Dirichlet Distribution. Master's Thesis, Queens University, Kingston, ON, Canada, 2016.
23. Robin, K.S. A generalization of the Dirichlet distribution. *J. Stat. Softw.* **2010**, *33*, 1–18.
24. Bilodeau, M.; Brenner, D. *Theory of Multivariate Statistics*; Springer: New York, NY, USA, 1999.
25. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions*; Dover Publications: Washington, DC, USA, 1964.
26. Bishop, C. Pattern Recognition and Machine Learning. *J. Electron. Imaging* **2006**, *4*, 049901. [CrossRef]
27. Zhanyu, M.; Pravin, K.R.; Jalil, T.; Markus, F.; Arne, L. Bayesian estimation of Dirichlet mixture model with variational inference. *Pattern Recognit.* **2014**, *47*, 3143–3157. [CrossRef]
28. Ma, Z. Bayesian estimation of the Dirichlet distribution with expectation propagation. In Proceedings of the 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 27–31 August 2012; pp. 689–693.
29. Zhuang, L.; Xu, A.; Wang, X.L. A prognostic driven predictive maintenance framework based on Bayesian deep learning. *Reliab. Eng. Syst. Saf.* **2023**, *234*, 109181. [CrossRef]
30. Devroye, L. *Non-Uniform Random Variate Generation*; Springer Science+Business Media: New York, NY, USA, 1986.
31. Keeping, E.S. *Introduction to Statistical Inference*; Dover Publications: New York, NY, USA, 1995; pp. 126–127.
32. Hines,W.W.; Montgomery, D.C.; Goldsman, D.M.; Borror, C.M. *Probability and Statistics in Engineering*, 4th ed.; John Wiley Sons, Inc.: Hoboken, NJ, USA, 2008; pp. 222–225.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.