mathematics

MDPI

Article

Compositional Data Modeling through Dirichlet Innovations

Seitebaleng Makgai [1,†], Andriette Bekker [1,†] and Mohammad Arashi [1,2,*]

[1] Department of Statistics, University of Pretoria, Pretoria 0028, South Africa; seite.makgai@up.ac.za (S.M.); andriette.bekker@up.ac.za (A.B.)
[2] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran
* Correspondence: arashi@um.ac.ir
† These authors contributed equally to this work.

Abstract: The Dirichlet distribution is a well-known candidate in modeling compositional data sets. However, in the presence of outliers, the Dirichlet distribution fails to model such data sets, making other model extensions necessary. In this paper, the Kummer–Dirichlet distribution and the gamma distribution are coupled, using the beta-generating technique. This development results in the proposal of the Kummer–Dirichlet gamma distribution, which presents greater flexibility in modeling compositional data sets. Some general properties, such as the probability density functions and the moments are presented for this new candidate. The method of maximum likelihood is applied in the estimation of the parameters. The usefulness of this model is demonstrated through the application of synthetic and real data sets, where outliers are present.

Keywords: beta function; compositional data; Dirichlet distribution; gamma distribution; Kummer–Dirichlet; outliers

Citation: Makgai, S.; Bekker, A.; Arashi, M. Compositional Data Modeling through Dirichlet Innovations. *Mathematics* **2021**, *9*, 2477. https://doi.org/10.3390/math9192477

Academic Editor: Tatjana von Rosen

Received: 29 July 2021
Accepted: 27 September 2021
Published: 3 October 2021

1. Introduction

Compositional data sets have played a valuable role in the medical, genetics and biological sciences due to the relative information conveyed through proportions, probabilities and percentages, as stated by [1]. Reference [1] describes the sample space of a compositional data set to be on a simplex, where the sum of all data points equals one or some whole number.

The most popular distribution that is well-known in modeling compositional data sets is the Dirichlet distribution (see for example [2]). Literature contains varying generalizations of the Dirichlet distribution that have been well studied in the application of various compositional data sets (see for example [3–8]). Other generalizations that are studied in the literature are part of the Liouville distribution as described in [9–11]. In Bayesian statistics, the Dirichlet distribution is known as a conjugate prior of the multinomial distribution and it is best used in estimating categorical distributions.

An extension of the Dirichlet distribution, known as the Dirichlet-generated class of distributions, has recently been introduced and developed by [12]. This extension served as a flexible alternative to the well-known Dirichlet and generalized Dirichlet distributions, where its aim is to address the limitations that the Dirichlet distribution may pose when modeling certain compositional data sets. Consider a compositional data set, where diagnostic probabilities of a sample of 15 students are assigned by clinicians. The background of this data set is further explained in Section 6. Figure 1 gives a scatterplot of the probabilities and illustrates the fit of the Dirichlet distribution (bivariate case) to this data set. Figure 1 illustrates an opportunity where the fit of the Dirichlet distribution could be improved upon.

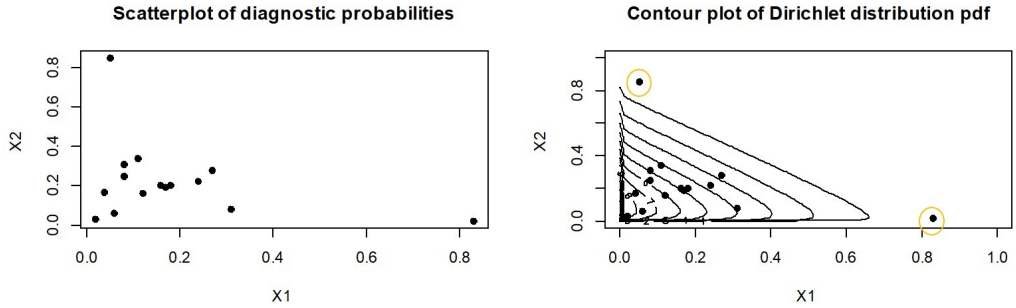

Figure 1. Plots of the data and the Dirichlet distribution on the diagnostic probabilities data set.

In [12], the beta-generating construction technique (pioneered and developed by [13]) is implemented to improve the fit of the Dirichlet distribution. The technique was an evolution from the univariate framework described below into a multivariate setting:

$$H(x) = \int_0^{G(x)} f(y)dy, \tag{1}$$

with the probability density function (pdf)

$$h(x) = f(G(x))g(x), \tag{2}$$

where $G(\cdot)$ is a continuous cumulative distribution function (cdf) and $f(\cdot)$ is the pdf of a random variable with support $[0, 1]$. By introducing extra parameters in $f(\cdot)$ and $G(\cdot)$, the resulting distribution provides greater flexibility in adapting modality and skewness.

Motivated by (1), from a multivariate viewpoint, in the methodology of [12], a new distribution $H(x_1, ..., x_p)$ for a random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$, $x_i > 0, i = 1, 2, \ldots, p$, is constructed by nesting the cdf of a baseline distributions $G_i(x_i)$ within the pdf of the generator distribution:

$$H(x_1, ..., x_p) = \frac{1}{B(\boldsymbol{\alpha})} \int_0^{G_1(x_1)} \cdots \int_0^{G_p(x_p)} \prod_{i=1}^{p} y_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{p} y_i\right)^{\alpha_{p+1} - 1} \boldsymbol{dy}, \tag{3}$$

with the pdf

$$h(x_1, ..., x_p) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{p} g_i(x_i) G_i(x_i)^{\alpha_i - 1} \left(1 - \sum_{i=1}^{p} G_i(x_i)\right)^{\alpha_{p+1} - 1}, \tag{4}$$

for $0 < y_i < 1$, $\sum_{i=1}^{p} y_i < 1$, $0 < G_i(x_i) < 1$ and where $B(\boldsymbol{\alpha})$ is the multivariate beta function. Here $\sum_{i=1}^{p} G_i(x_i) < 1$ and $g_i(x_i)$ and $G_i(x_i)$ are the pdf and cdf of the baseline distributions, respectively. The authors [12] developed the Dirichlet-gamma distribution, where in this case, the gamma distribution is taken as the baseline distribution $G_i(x_i)$, $i = 1, 2, \cdots, p$, and the Dirichlet distribution is taken as the generator distribution.

In the univariate case, the Kummer-beta distribution is seen as an extension of the beta distribution (see the studies of [14–16]), it then follows that the multivariate Kummer-beta (refer as to Kummer–Dirichlet hereafter) distribution is also considered as an extension of the Dirichlet distribution (see [17]). Authors such as [14–16,18] have applied the generating technique to the Kummer-beta distribution, by coupling the cdf of different baseline distributions with the pdf of the Kummer-beta distribution. The development of generated distributions using the Kummer-beta distribution, has introduced distributions that add more flexibility in modeling data sets that are in the $(0, 1)$ domain (see [19] for an example).

In this paper, we propose a general multivariate construction methodology using the Kummer–Dirichlet (KD) pdf as the generator. This KD-generated class serves as a good alternative to the Dirichlet distribution for the statistical representation of specific

proportional data. This class can be viewed as an evolution from the univariate framework into a multivariate setting as described in (3) but with the aim of offering more flexibility in modeling compositional data sets.

Thus, we introduce the KD distribution as the generating distribution, and a new class is proposed, with the following cdf

$$H(x_1,...,x_p) = C\int_0^{G_1(x_1)}\cdots\int_0^{G_p(x_p)}\prod_{i=1}^{p} y_i^{\alpha_i-1}\left(1-\sum_{i=1}^{p} y_i\right)^{\alpha_{p+1}-1} exp\left(-\lambda\sum_{i=1}^{p} y_i\right)\boldsymbol{dy}, \quad (5)$$

with $\alpha_i > 0$ for $i = 1,2,\cdots,p+1$, $-\infty < \lambda < \infty$, C as the normalizing constant, $0 < y_i < 1$, $\sum_{i=1}^{p} y_i < 1$, $\boldsymbol{y} = (y_1, y_2, \cdots, y_p)$ and $G_i(x_i)$, $i = 1,2,\cdots,p$, as the cdfs of a baseline distribution with $\sum_{i=1}^{p} G_i(x_i) < 1$. Distributions with cdf (5) and normalizing constant (9) shall be referred to as Kummer–Dirichlet generated distributions, where $G_i(x_i)$, $i = 1,2,\cdots,p$, are the cdfs of a baseline distribution.

The contribution of this construction (5) highlights the importance of developing distributions that can improve the modeling of extreme observations in compositional data sets, where the Dirichlet might not be suitable or at a shortfall, as illustrated in Figure 1. For such cases and others that may arise, we propose a model with cdf (5). Thus, this novel study contributes to multivariate distribution theory from the following aspects:

1. The well-known beta-generator in the univariate case is extended to the Kummer–Dirichlet in the multivariate case.
2. A technique is proposed to construct multivariate distributions that combines a baseline distribution with a multivariate generator and evolves generating a plethora of possibilities of results.
3. We proposed a multivariate distribution that can be used for modeling compositional data with outliers.
4. Mathematical techniques are developed to derive the moment generating function of multivariate distributions.

The following showcases the organization of our contribution; in Section 2, the building blocks for the KD generator distribution, such as the normalizing constant of the pdf that corresponds to (5) is derived. In Section 3, the KD-gamma distribution is introduced, where we provide some technical results to derive the moments. In Section 4, the usefulness of the KD-Gamma distribution, as compared to the Dirichlet-gamma distribution, is seen through the application of a synthetic data analysis. Two real data sets, where outliers are present, are analyzed in Section 5. Finally, some conclusions are given in Section 6. Proof of the main results are put in the Appendix A.

2. Building Blocks of the Kummer–Dirichlet Distribution

The building blocks and notations necessary in the construction of distributions with cdf (5) are presented in this section. Since the Dirichlet distribution is an important building block, it is known that a random vector $\boldsymbol{Y} = (Y_1,\ldots,Y_p) \in \mathcal{R}^p$ is said to be Dirichlet (or standard Dirichlet) distributed with parameters $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_p;\ \alpha_{p+1})$ for $\alpha_i > 0, i = 1,...,p+1$, $p \geq 2$, if its pdf is given by

$$f(\boldsymbol{y}) = C_1(\boldsymbol{\alpha}) y_1^{\alpha_1-1}\cdots y_p^{\alpha_p-1}\left(1-\sum_{i=1}^{p} y_i\right)^{\alpha_{p+1}-1}. \quad (6)$$

From (6), one can denote $Y_{p+1} = 1-\sum_{i=1}^{p} Y_i$ and let $\boldsymbol{Y}' = (Y_1,\ldots,Y_p;Y_{p+1}) = (\boldsymbol{Y};Y_{p+1})$. The random vectors $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ can be defined on $\boldsymbol{\Omega}_p$ and $\boldsymbol{\mathcal{S}}_{p+1}$, respectively, where

$$\boldsymbol{\Omega}_p = \left\{(y_1,\ldots,y_p) \in \mathcal{R}^p : \sum_{i=1}^{p} y_i < 1, y_i > 0, i = 1,...,p\right\}$$

and

$$\boldsymbol{\mathcal{S}}_{p+1} = \left\{ (y_1, \ldots, y_{p+1}) \in \mathcal{R}^{p+1} : \sum_{i=1}^{p+1} y_i = 1, y_i > 0, i = 1, ..., p+1 \right\},$$

for $p \geq 2$. The constant $C_1(\boldsymbol{\alpha})$ in (6) is given as

$$C_1^{-1}(\boldsymbol{\alpha}) = \int_{\boldsymbol{\Omega}_p} \prod_{i=1}^{p+1} y_i^{\alpha_i - 1} \boldsymbol{dy} = \frac{\prod_{i=1}^{p+1} \Gamma(\alpha_i)}{\Gamma(\alpha_+)} = B(\boldsymbol{\alpha}), \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function. Now using the Kummer-beta distribution (see [14]) as foundation building blocks, it follows that a random vector $\boldsymbol{Y}^{'} = (Y_1, \ldots, Y_p; Y_{p+1}) = (\boldsymbol{Y}; Y_{p+1})$ is said to be multivariate Kummer–Dirichlet distributed with parameters $(\boldsymbol{\alpha}, \lambda) = (\alpha_1, \cdots, \alpha_p;\ \alpha_{p+1}, \lambda)$ for $\alpha_i > 0, i = 1, ..., p+1$, $p \geq 2$ and $-\infty < \lambda < \infty$, if its pdf is given by

$$f(\boldsymbol{y}) = C_2(\boldsymbol{\alpha}, \lambda)\, y_1^{\alpha_1 - 1} \cdots y_p^{\alpha_p - 1} \left(1 - \sum_{i=1}^{p} y_i\right)^{\alpha_{p+1} - 1} exp\left(-\lambda \sum_{i=1}^{p} y_i\right), \tag{8}$$

where $y_i > 0$ and $\sum_{i=1}^{p} y_i < 1$ for $i = 1, ..., p$. The following theorem gives the derivation of the normalizing constant $C_2(\boldsymbol{\alpha}, \lambda)$.

Theorem 1. *In the general case of $p \geq 2$, the normalizing constant $C_2(\boldsymbol{\alpha}, \lambda)$ in pdf (8) is given by*

$$\begin{aligned} \frac{1}{C_2(\boldsymbol{\alpha}, \lambda)} &= \frac{\prod_{i=1}^{p} \Gamma(\alpha_i)\Gamma(\alpha_{p+1})}{\Gamma(\sum_{i=1}^{p} \alpha_i + \Gamma(\alpha_{p+1}))} \sum_{m_1, \cdots, m_p \geq 0}^{\infty} \left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!} \right) \frac{\prod_{i=1}^{p} (\alpha_i)_{m_i}}{(\sum_{i=1}^{p} \alpha_i + \alpha_{p+1})_{\sum_{i=1}^{p} m_i}} \\ &= \frac{\prod_{i=1}^{p} \Gamma(\alpha_i)\Gamma(\alpha_{p+1})}{\Gamma(\sum_{i=1}^{p} \alpha_i + \alpha_{p+1})}\, {}_1F_1\left(\sum_{i=1}^{p} \alpha_i; \sum_{i=1}^{p} \alpha_i + \alpha_{p+1}; -\lambda \right), \end{aligned} \tag{9}$$

where $\alpha_i > 0$ for $i = 1, 2, \cdots, p$, $-\infty < \lambda < \infty$, $(\alpha)_n$ denotes the Pochhammer function $(\alpha)_n = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}$ and ${}_1F_1(.;.;.)$ is the confluent hypergeometric function.

For the proof, refer to Appendix A.

2.1. Kummer–Dirichlet Generator

In this section, we give the definition of KD generated distribution with some technicalities.

Definition 1. *A random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$ is said to follow a Kummer–Dirichlet generated distribution, if its cdf is given by (5) and has pdf*

$$h(\boldsymbol{x}) = h(x_1, ..., x_p) = C_2(\boldsymbol{\alpha}, \lambda) \left(1 - \sum_{i=1}^{p} G_i(x_i)\right)^{\alpha_{p+1} - 1} \prod_{i=1}^{p} g_i(x_i) G_i^{\alpha_i - 1}(x_i) exp(-\lambda G_i(x_i)), \tag{10}$$

where $C_2(\boldsymbol{\alpha}, \lambda)$ is the normalizing constant (9), and where shape parameters $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_{p+1})$ are all > 0 , $-\infty < \lambda < \infty$, $g_i(x_i)$ and $G_i(x_i)$ $i = 1, 2, \cdots, p$ as the pdfs and cdfs, respectively, of the baseline distribution for $\sum_{i=1}^{p} G_i(x_i) < 1$. The random vector is then denoted as $\boldsymbol{X} \sim KDG(\boldsymbol{\psi})$, where $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \lambda, \boldsymbol{\rho})$ with $\boldsymbol{\rho}$ as the parameters of the baseline distribution.

2.1.1. Special Cases

From cdf (5) and pdf (10), stem two classes of distributions as special cases of the Kummer–Dirichlet generated distribution.

- Class of Dirichlet-generated distributions: When $\lambda = 0$, the pdf (10) simplifies to the pdf of a Dirichlet-generated distribution, with baseline distribution $G(.)$ and beta-generated marginal distributions (see [12,13]).
- Class of Exponentiated Generalized-generated distributions: When $\lambda = 0$ and $\alpha_{p+1} = 1$, then the pdf (10) tends to the multivariate exponentiated-generalized distribution (this distribution is not yet introduced in literature), whose marginal distributions are exponentiated-generalized distribution (see [20]).

2.1.2. Expansions and Marginals of the Kummer–Dirichlet Generated Distributions

Expanding and re-writing the exponential term $exp(-\lambda G_i(x_i))$ in series form in (10), results in an infinite weighted sum of Dirichlet-generated distributions, where in this case, the pdf (10) is given by

$$\begin{aligned} h(\boldsymbol{x}) = h(x_1, ..., x_p) &= C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_1, \cdots, m_p \geq 0}^{\infty} \left(\prod_{j=1}^{p} \frac{(-\lambda)^{m_j}}{m_j!} \right) \left(1 - \sum_{i=1}^{p} G_i(x_i) \right)^{\alpha_{p+1}-1} \\ &\quad \times \prod_{i=1}^{p} g_i(x_i) G_i^{\alpha_i + m_i - 1}(x_i) \\ &= C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_1, \cdots, m_p \geq 0}^{\infty} w_{m_j} \left(1 - \sum_{i=1}^{p} G_i(x_i) \right)^{\alpha_{p+1}-1} \\ &\quad \times \prod_{i=1}^{p} g_i(x_i)\ G_i^{\alpha_i + m_i - 1}(x_i), \end{aligned} \tag{11}$$

where the coefficient $w_{m_j} = \left(\prod_{j=1}^{p} \frac{(-\lambda)^{m_j}}{m_j!} \right)$ can be considered as the weights for $m_j \geq 0$.

The binomial expansion in (11) where $\sum_{i=1}^{p} G_i(x_i) < 1$, can be expressed as

$$\left(1 - \sum_{i=1}^{p} G_i(x_i) \right)^{\alpha_{p+1}-1} = \sum_{k=0}^{\infty} (-1)^k \binom{\alpha_{p+1} - 1}{k} \left(G_1(x_1) + G_2(x_2) + \cdots + G_p(x_p) \right)^k$$

$$= \sum_{k=0}^{\infty} \sum_{\substack{v_1, \cdots, v_p \geq 0 \\ v_1 + \cdots + v_p = k}}^{k} (-1)^k \frac{k!}{v_1! \cdots v_p!} \binom{\alpha_{p+1} - 1}{k} G_1^{v_1}(x_1) \cdots G_p^{v_p}(x_p). \tag{12}$$

It follows from (11) and (12) that the pdf of the Kummer–Dirichlet generated distribution can also be expressed as a linear combination of exponentiated distributions that were introduced by [21] and then expanded by [20,22,23], where the Weibull distribution was taken as the baseline distribution. Hence,

$$h(x_1, ..., x_p) = C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_j, v_j, k \geq 0}^{\infty} w_{m_j, v_j, k} \prod_{i=1}^{p} g_i(x_i)\ G_i^{\alpha_i + m_i + v_i - 1}(x_i), \tag{13}$$

where $j = 1, 2, \cdots, p$ and the coefficient $w_{m_j, v_j, k}$ given as

$$w_{m_j, v_j, k} = \left(\prod_{j=1}^{p} \frac{(-\lambda)^{m_j}}{m_j!} \right) (-1)^k \frac{k!}{v_1! \cdots v_p!} \binom{\alpha_{p+1} - 1}{k}. \tag{14}$$

The marginal pdfs of X_i for $i = 1, 2, \cdots, p$ if $\boldsymbol{X} \sim KDG(\boldsymbol{\psi})$, $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \lambda, \boldsymbol{\rho})$ (see (10)), is given as

$$\begin{aligned} h_i(x_i) &= C_i(\boldsymbol{\alpha}, \lambda)\, g_i(x_i) G_i^{\alpha_i - 1}(x_i)(1 - G_i(x_i))^{\sum_{j=1, \neq i}^{p+1} \alpha_j - 1} exp(-\lambda G_i(x_i)) \\ &\quad \times {}_1F_1\left(\alpha_{p+1}; \sum_{j=i+1}^{p} \alpha_j + \alpha_{p+1}; \lambda(1 - G_i(x_i))\right), \end{aligned} \tag{15}$$

where $C_i(\boldsymbol{\alpha}, \lambda)$ is the normalizing constant of the marginal distribution, for $i = 1, 2, \cdots, p$, $g_i(.)$ and $G_i(.)$ for $i = 1, 2, \cdots, p$ as the pdfs and cdfs, respectively, of the baseline distribution.

3. The Kummer–Dirichlet Gamma Distribution

In this section, we focus on the gamma distribution as the chosen baseline distribution. The gamma distribution, which belongs to the exponential class, is a flexible distribution model with a shape parameter, that may offer a good fit to a variety of different data sets [24]. The cdf and pdf of the gamma distribution with shape parameter $\delta > 0$ and scale parameter $\theta > 0$ are given as

$$G(x; \delta, \theta) = \frac{\gamma(\delta, \frac{x}{\theta})}{\Gamma(\delta)} \tag{16}$$

and pdf

$$g(x; \delta, \theta) = \frac{1}{\theta^{\delta}\Gamma(\delta)} x^{\delta} e^{-\frac{x}{\theta}}, \tag{17}$$

where $\gamma(\delta, \frac{x}{\theta})$ is the incomplete gamma function $\int_0^x t^{\delta - 1} e^{-t} dt$.

Thus, here, we explore the impact of the gamma distribution as the considered baseline distribution, where the cdf and pdf of the baseline distribution is given by (16) and pdf (17), respectively. In this case, $G_i(.)$ for $i = 1, 2, \cdots, p$ are the cdfs of the gamma distribution with shape and scale parameters $\boldsymbol{\rho} = (\boldsymbol{\delta}, \boldsymbol{\theta})$ for $\delta_i > 0, \theta_i > min(x_i)$, $i = 1, 2, \cdots, p$, we denote random vector $\boldsymbol{X} \sim KDGa(\boldsymbol{\psi})$ as Kummer–Dirichlet gamma (KDGa) distributed where $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \lambda, \boldsymbol{\delta}, \boldsymbol{\theta})$.

Figures 2–4 illustrate the effect of the parameters $(\alpha_1, \alpha_2, \alpha_3, \delta_1, \theta_1, \delta_2, \theta_2, \lambda)$ of the pdf (10). It is observed in Figure 2 that parameters $(\alpha_1, \alpha_2, \alpha_3)$ illustrate the influence or "weight" of each random variable X_i, in this case $i = 1, 2$. From Figure 2, it is observed that larger values of α_1 leads to skewness and heavier tails for random variable X_1. Symmetry is observed in the first row of Figure 2 when $(\alpha_1, \alpha_2, \alpha_3) = (2, 2, 2)$. The parameters $(\delta_1, \theta_1, \delta_2, \theta_2)$ influence the shape, peakness and the scale of the pdf (10). It is observed in Figure 3 that smaller values of δ_i, $i = 1, 2$ results in the pdf (10) concentrated on a smaller scale, while larger values of δ_i, $i = 1, 2$ results in the pdf (10) spread across a bigger scale of values. It is observed in Figure 4 that λ influences the tails, peakness and narrowness of the pdf (10). It is observed in the first row of Figure 4 that smaller values of λ results in heavier tails.

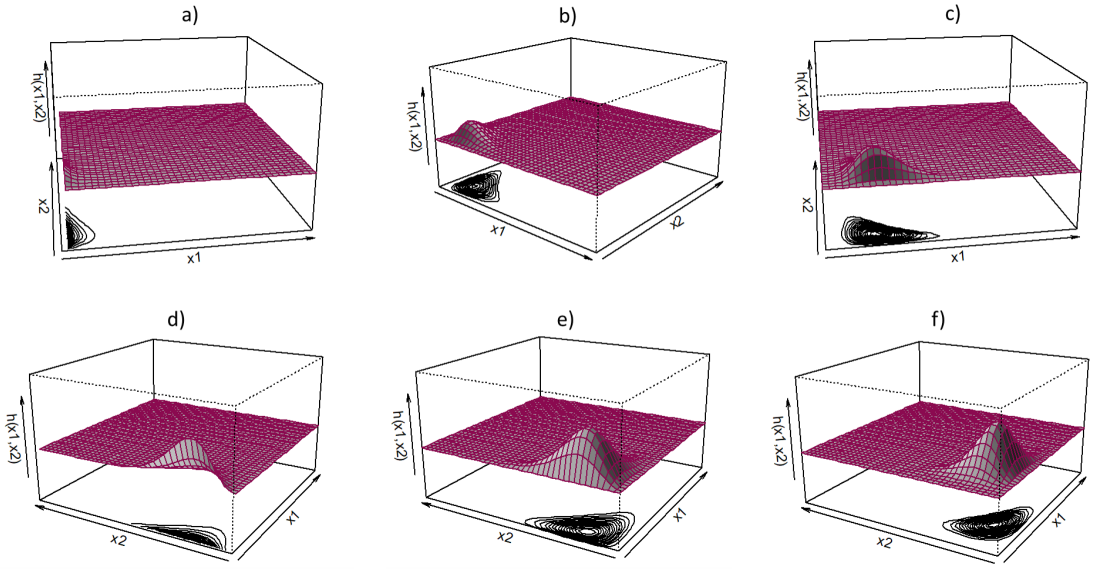

Figure 2. Example pdfs and contour plots for (10) for $(\alpha_1, \alpha_2, \alpha_3, \delta_1, \theta_1, \delta_2, \theta_2, \lambda)$ when (**a**) (0.1,2,2,2,2,2,2,2), (**b**) (2,2,2,2,2,2,2,2), (**c**) (4,2,2,2,2,2,2,2), (**d**) (0.1,2,4,2,1.5,2,1.5,−2), (**e**) (1,2,4,2,1.5,2,1.5,−2) and (**f**) (4,2,4,2,1.5,2,1.5,−2).

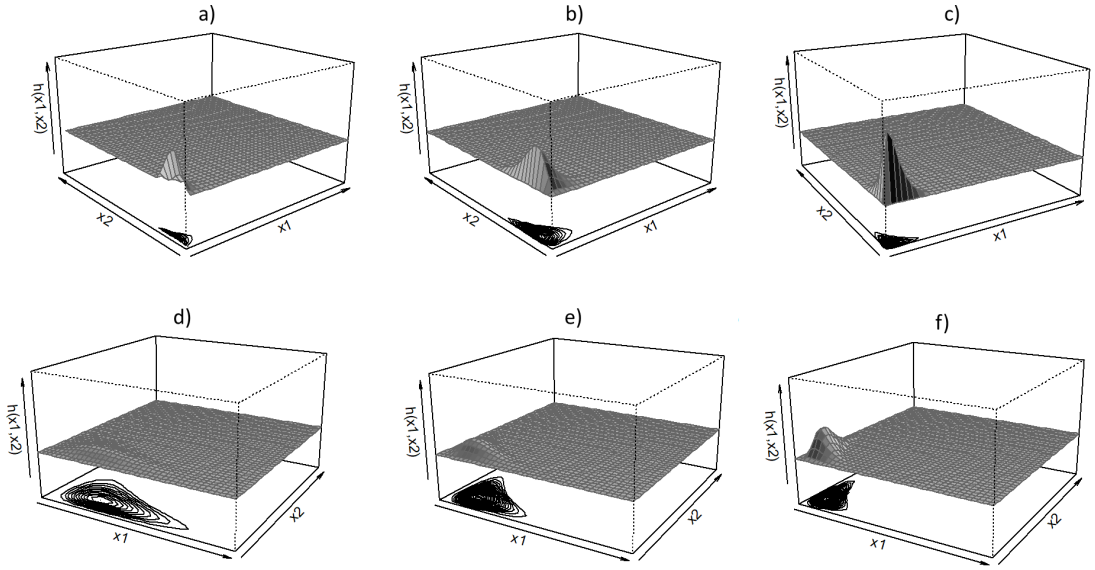

Figure 3. Example pdfs and contour plots of (10) for various values of $(\alpha_1, \alpha_2, \alpha_3, \delta_1, \theta_1, \delta_2, \theta_2, \lambda)$ when (**a**) (2,2,2,0.5,2,2,2,2), (**b**) (2,2,2,1,2,2,2,2), (**c**) (2,2,2,1,2,1,2,2), (**d**) (2,2,2,1.8,0.8,2,2,2), (**e**) (2,2,2,1.8,1.5,2,2,2), (**f**) (2,2,2,1.8,3,2,2,2).

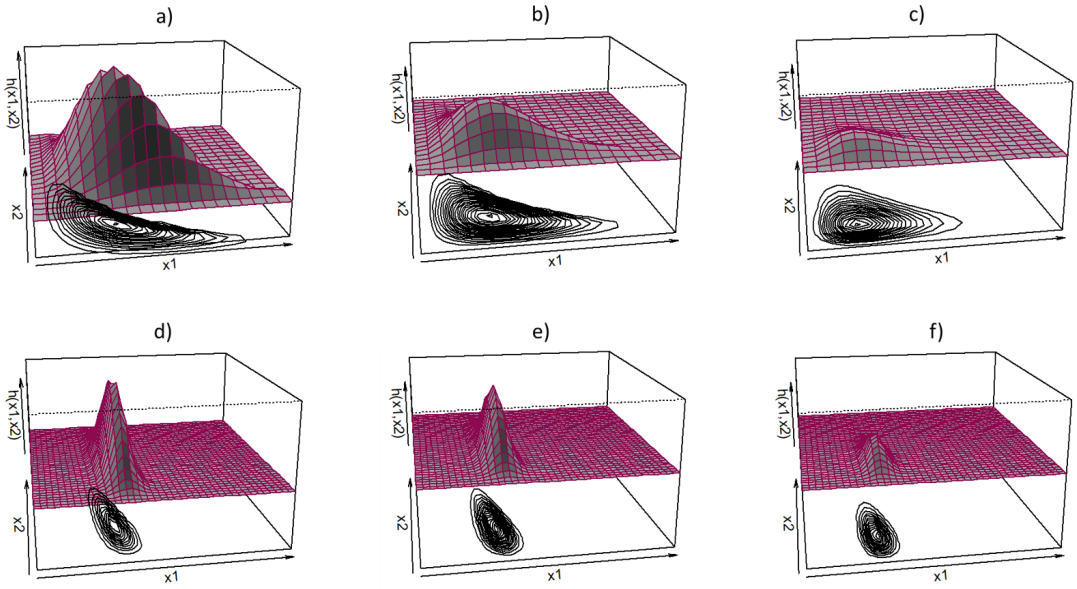

Figure 4. Example pdfs and contour plots of (10) for various values of $(\alpha_1, \alpha_2, \alpha_3, \delta_1, \theta_1, \delta_2, \theta_2, \lambda)$ when (**a**) (2,2,2,2,2,2,2,−4), (**b**) (2,2,2,2,2,2,2,0), (**c**) (2,2,2,2,2,2,2,4), (**d**) (4,2,8,12,8,2,0.5,−8), (**e**) (4,2,8,12,8,2,0.5,−4), (**f**) (4,2,8,12,8,2,0.5,4).

Moment Generating Function of the KDG

In this section, the moment generating function (mgf) and product moments of random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_p) \sim KDGa(\boldsymbol{\psi})$, where $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \lambda, \boldsymbol{\delta}, \boldsymbol{\theta})$ are derived.

Theorem 2. *The mgf of random vector* $\boldsymbol{X} \sim KDGa(\boldsymbol{\alpha}, \lambda, \boldsymbol{\delta}, \boldsymbol{\theta})$ *is given by*

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_j, v_j, k \geq 0}^{\infty} w_{m_j, v_j, k} \times \prod_{i=1}^{p} \frac{1}{\theta_i^{\delta_i} (\frac{1}{\theta_i} - t_i)^{\delta_i}} \cdot \frac{1}{\alpha_i + m_i + v_i}, \tag{18}$$

where $\boldsymbol{t} = (t_1, \ldots, t_p)$, $C_2(\boldsymbol{\alpha}, \lambda)$ *is the normalizing constant (9), shape parameters* $\boldsymbol{\alpha} = \alpha_1, \cdots, \alpha_p$, $\alpha_{p+1} > 0$, $w_{m_j, v_j, k}$ *as the coefficient given by (14) for* $j = 1, 2, \cdots, p$, $-\infty < \lambda < \infty$, *shape parameter* $\delta_i > 0$ *and scale parameter* $\theta_i = min(x_i) > 0$ *for* $i = 1, 2, \cdots, p$.

For the proof, refer to Appendix A.

Theorem 3. *Let* n_i, $i = 1, \ldots, p$ *be positive integer values. Then, the product moments of* $\boldsymbol{X} \sim KDGa(\boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\delta}, \lambda)$ *is expressed in the following form*

$$\begin{aligned} \mathcal{E} &= E\left[\prod_{i=1}^{p} X_i^{n_i}\right] \\ &= C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_j, v_j, k \geq 0}^{\infty} w_{m_j, v_j, k} \prod_{i=1}^{p} \frac{\theta_i^{n_i} \Gamma(n_i + \delta_i)}{\Gamma(\delta_i)(\alpha_i + m_i + v_i)}, \end{aligned} \tag{19}$$

where $C_2(\boldsymbol{\alpha}, \lambda)$ *is the normalizing constant (9), shape parameters* $\boldsymbol{\alpha} = \alpha_1, \cdots, \alpha_p, \alpha_{p+1} > 0$, $w_{m_j, v_j, k}$ *as the coefficient given by (14) for* $j = 1, 2, \cdots, p$, $-\infty < \lambda < \infty$, *shape parameter* $\delta_i > 0$ *and scale parameter* $\theta_i = min(x_i) > 0$ *for* $i = 1, 2, \cdots, p$.

For the proof, refer to Appendix A.

For the illustration section and ease of reader, the moments for the bivariate case ($p = 2$) of the Kummer=-Dirichlet gamma distribution is given as

$$E[X_1^r X_2^s] = C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_1,m_2\geq 0}^{\infty} \sum_{k=0}^{\infty} \sum_{v_1,v_2\geq 0}^{k} \frac{(-\lambda)^{m_1+m_2}}{m_1!m_2!} \frac{(-1)^k k!}{v_1!v_2!} \begin{pmatrix} \alpha_3 - 1 \\ k \end{pmatrix} \times \frac{\theta_1^2\theta_2^2\Gamma(\delta_1 + r)\Gamma(\delta_2 + s)}{\Gamma(\delta_1)\Gamma(\delta_2)(\alpha_1 + m_1 + v_1)(\alpha_2 + m_2 + v_2)}, \tag{20}$$

using the result of (15) and (19).

4. Synthetic Data Analysis

In this section, the performance of the Kummer–Dirichlet gamma and Dirichlet-gamma distributions are analyzed to illustrate the model capabilities for a synthetic data set.

4.1. Study 1

In the first simulation study, an artificial data set is generated via a specified seed value and through Weibull random variates using Algorithm 1. For this synthetic data, the Weibull random variates w_i, $i = 1, 2, 3$ are generated using R; assuming that the random variable W is Weibull distributed [24], if W has cdf:

$$G(w) = 1 - \exp\left(-\left(\frac{w}{\xi}\right)^{\nu}\right) \tag{21}$$

and pdf

$$g(w) = \frac{\nu}{\xi^{\nu}} w^{\nu-1} \exp\left(-\left(\frac{w}{\xi}\right)^{\nu}\right), \tag{22}$$

where $w \geq 0$ with shape parameter $\nu > 0$ and scale parameter $\xi > 0$. The construction of this synthetic data set, results in a compositional data set with a negative correlation. The seed for generating Weibull random variates is set at 7, with initial parameter values $W_1 \sim Wei(1.5, 1.5)$, $W_2 \sim Wei(4, 2)$ and $W_3 \sim Wei(4, 1)$.

Algorithm 1 Synthetic data generation using the Weibull distribution.

Step 1. Generate 100 random variates $W_i \sim Wei(\xi_i, \nu_i)$ for $i = 1, 2, 3$.

Step 2. Define random variables $Y = (Y_1, Y_2, Y_3)$, where $Y_i = \frac{W_i}{\sum_{i=1}^{3} W_i}$, $i = 1, 2, 3$ and simulate a synthetic data set $y = (y_1, y_2, y_3)$.

To measure the fit of the Kummer–Dirichlet gamma vs. the Dirichlet-gamma distribution, a ratio of Kolmogorov–Smirnov (KS) distance measures are calculated over a number of simulations as given in Algorithms 2 and 3. The following Algorithm 2 gives the steps used to assess the competence of the models. This ratio of KS measures describes a model testing technique developed by [12], called the empirical estimator of the cdf of a multivariate distribution. The technique compares the empirical cdfs of the observed and simulated datasets. The advantage of this technique is that one can also use the empirical cdfs to rank the simulated data. Ranking data makes it possible to calculate more accurate distances between the observed data points and the simulated points. More details regarding this technique of ranking simulated data in order to calculate the optimal distance between data points are available in [12].

This technique is used here to analyze the performances of the Dirichlet-gamma (DGa) and the Kummer–Dirichlet gamma ($KDGa$) distributions. In this technique, the empirical cdf of the generated data set and the cdf of the analyzed distributions are used in calculating

a ratio of Kolmogorov–Smirnov (KS) distance measures, where the model with the smallest KS measure is considered as the most suitable amongst the two.

Algorithm 2 Computing the KS ratio measure.

Step 1. Calculate the empirical cdf $\widehat{F(\mathbf{x})} = P\left(X_1^{(i)} \leq x_1, X_2^{(i)} \leq x_2, \cdots, X_p^{(i)} \leq x_p\right)$ for data points $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$, where $i = (1, \ldots, n)$ and n as the sample size.
Step 2. Obtain the estimates of the parameters of the two models, *KDGa* and *DGa* distributions.
Step 3. Simulate two data sets of sizes $d > n$ using the parameter estimates obtained in Step 2; $x^*_{KDGa} = \left(x_1^*, x_2^*, \cdots, x_p^*\right)$ and $x^*_{DGa} = \left(x_1^*, x_2^*, \cdots, x_p^*\right)$.
Step 4. For each generated data set x^*_{KDGa} and x^*_{DGa}, calculate the empirical cdfs $\widehat{F(\mathbf{x}^*)} = P\left(X_1^* \leq x_1, X_2^* \leq x_2, \cdots, X_p^* \leq x_p\right)$.
Step 5. Compute the KS distance measures between $\widehat{F(\mathbf{x})}$ and $\widehat{F(\mathbf{x}^*)}$ as computed in Steps 2–4, where in this case, $KS = \max\left|\widehat{F(\mathbf{x}^*)} - \widehat{F(\mathbf{x})}\right|$.
Step 6. Repeat Steps 2–5 m times, and compute the average of the KS measures.
Step 7. Compare the average KS measures of the *KDGa* to the average KS measure of the *DGa* using the ratio $\frac{KS\ of KDGa}{KS\ of\ DGa}$.

Algorithm 3 Computing the KS ratio measure using Weibull.

Step 1. Generate a synthetic non-Dirichlet data set using the Weibull random variates as computed before Algorithm 1 and take the data set as the observed data.
Step 2. Using this observed data set, obtain parameter estimates for the Kummer–Dirichlet gamma and Dirichlet-gamma distributions as the proposed models.
Step 3. From the obtained parameter estimates simulate data sets of sizes $d = 50$ and 100. Calculate the empirical cdfs $\widehat{F(\mathbf{x}^*)}$ for each simulation, as seen in Step 4 of Algorithm 2.
Step 4. Calculate the KS measures between the empirical cdf $\widehat{F(\mathbf{x})}$ and the cdfs of the two competing models $\widehat{F(\mathbf{x}^*)}$, for each group.
Step 5. Repeat Steps 3 and 4, 50 times and compute the average KS measure for the two models.
Step 6. Represent the KS measure of the *KDGa* and *DGa* as a ratio $\frac{KS\ of KDGa}{KS\ of DGa}$ for each simulated group of $d = 50$ and 100.

It is observed in Figure 5 that the Kummer–Dirichlet gamma distribution adds additional coverage to the generated artificial non-Dirichlet data set. A KS ratio of 0.89:1 illustrates that the KS distance of Kummer–Dirichlet gamma is 11% less of the KS distance measure of the Dirichlet-gamma distribution. This ratio indicates that through the numerous simulations, the KS measure of the Kummer–Dirichlet gamma distribution was observed to be smaller than the KS measure of the Dirichlet-gamma distribution.

4.2. Study 2

In this simulation, the Expected-Modification (EM) algorithm is used to estimate the parameters for generated samples of sizes 50 and 100 of the Kummer–Dirichlet gamma distribution. The EM algorithm is considered here, since the pdf of a Kummer–Dirichlet generated distribution can be expressed as a mixture of its special cases. The EM algorithm consists essentially of two main steps; the Expectation and Modification steps, with the main aim of maximizing the log-likelihood function $ll(\boldsymbol{\psi})$ of the observed data with respect to the unknown vector of parameters $\boldsymbol{\psi}$. It is summarized as follows:

Step 1: *The E-step*: In this step, the missing data Z are computed.

Step 2: *The M-step:* In this step, obtain the parameter estimates that maximizes $K = E[ln\, h(\boldsymbol{X}|\boldsymbol{\psi})|Z]$, where $ln(\, h(\boldsymbol{X}|\boldsymbol{\psi}))$ is the log-likelihood function and $h(\boldsymbol{X}|\boldsymbol{\psi})$ is the pdf (10).

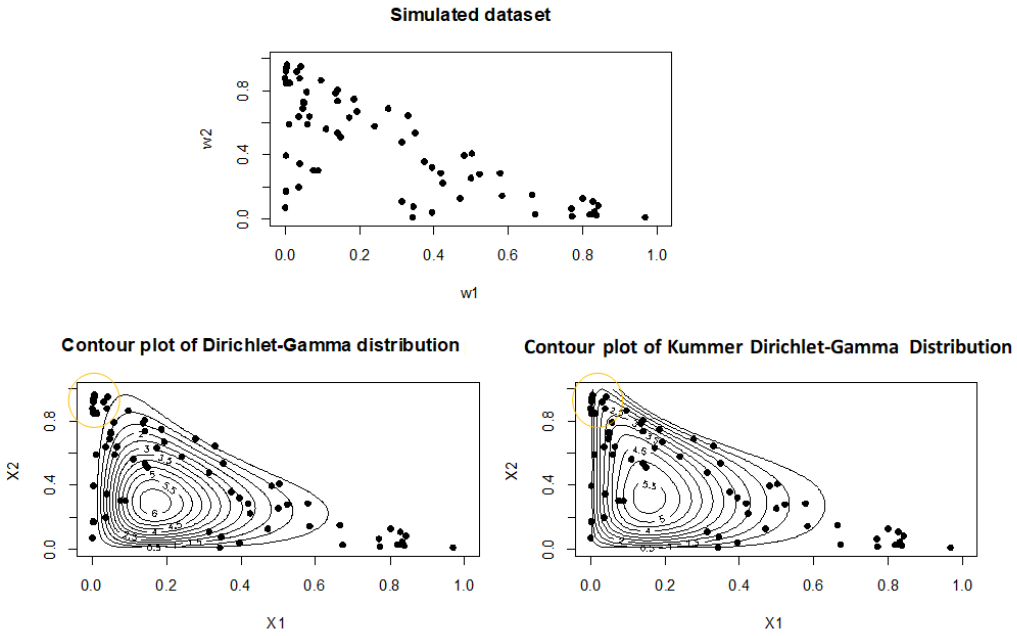

Figure 5. Performance of the Kummer–Dirichlet gamma and Dirichlet-gamma distributions on the synthetic compositional data set generated using Algorithm 3.

In the bivariate case, let Q be the *observed* data (generated through Algorithm 4), let Z be the missing data and let $X^* = (Q, Z)$ be the *complete* data set. In the case where samples of sizes $n = 50$ and $n = 100$ are drawn, let

$$ll(\boldsymbol{\psi}) = \sum_{i=1}^{n} \log h(x_{1i}^*, x_{2i}^*; \boldsymbol{\psi})$$

be the log-likelihood function based on the complete data X^* with parameters $\boldsymbol{\psi} = (\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\theta}, \lambda)$.

Algorithm 4 Generation of observed data for the EM algorithm.

Step 1. Generate a random sample $U \sim Unif(0,1)$ of size 30.
Step 2. Generate a random sample of size 30 from the marginal distributions (15), where

$$C(\boldsymbol{\alpha}, \lambda) \int_0^{G_1(q_1)} y_1^{\alpha_1 - 1}(1 - y_1)^{\alpha_2 + \alpha_3 - 1} e^{-\lambda y_1}\, {}_1F_1(\alpha_3; \alpha_2 + \alpha_3; \lambda(1 - y_1)) dy_1 = U$$

$$C(\boldsymbol{\alpha}, \lambda) \int_0^{G_2(q_2)} y_2^{\alpha_2 - 1}(1 - y_2)^{\alpha_1 + \alpha_3 - 1} e^{-\lambda y_2}\, {}_1F_1(\alpha_3; \alpha_1 + \alpha_3; \lambda(1 - y_2)) dy_2 = U$$

Step 3. Use the *Unitroot* function in R software to solve for q_1 and q_2, where $G_i(q_i)$ is the cdf of the gamma distribution (16).
Step 4. Observe data $Q = (q_1, q_2)$.

To compute the missing data, let

$$Z_{i,j} = \begin{cases} 1 & \text{if } q_i \text{ is from class } j \\ 0 & \text{otherwise} \end{cases}, \tag{23}$$

for q_i, $i = 1, 2$.

Samples of sizes 50 and 100 are generated using 100 trials for each group of fixed parameters. Hence 100 MLE's of the model parameters (using the procedure in R package *optim*) is obtained. The mean, bias and mean square error (MSE)

$$\text{Bias} = \frac{1}{100}\sum_{k=1}^{100}\widehat{\psi}_k - \psi_{true} \quad \text{and} \quad \text{MSE} = \frac{1}{100}\sum_{k=1}^{100}\left(\widehat{\psi}_k - \psi_{true}\right)^2,$$

are calculated. In this case, $\widehat{\psi}_k$ denotes the ML estimate of ψ_{true} (chosen parameter values) at the *kth* replication.

Table 1 gives the results of simulation study 2, for chosen parameter values $\boldsymbol{\psi} = (\alpha_1, \alpha_2, \alpha_3, \delta_1, \delta_2, \theta_1, \theta_2, \lambda) = (3.03, 7.92, 4.10, 1.96, 1.32, 0.36, 0.59, 1)$. The results in Table 1 illustrate that the mean, MSE and bias of the parameter estimates decreases for larger sample sizes (n). The length of the asymptotic confidence intervals also decrease for increasing sample size.

Table 1. Simulation results for sample size $n = 50$ and $\boldsymbol{\psi} = (\alpha_1, \alpha_2, \alpha_3, \delta_1, \delta_2, \theta_1, \theta_2, \lambda) = (3.03, 7.92, 4.10, 1.96, 1.32, 0.36, 0.59, 1)$.

$\boldsymbol{n = 50}$	$\widehat{\boldsymbol{\alpha}}_{\mathbf{1}}$	$\widehat{\boldsymbol{\alpha}}_{\mathbf{2}}$	$\widehat{\boldsymbol{\alpha}}_{\mathbf{3}}$	$\widehat{\boldsymbol{\delta}}_{\mathbf{1}}$	$\widehat{\boldsymbol{\delta}}_{\mathbf{2}}$	$\widehat{\boldsymbol{\theta}}_{\mathbf{1}}$	$\widehat{\boldsymbol{\theta}}_{\mathbf{2}}$	$\widehat{\boldsymbol{\lambda}}$
Mean	3.111	9.204	5.973	2.186	2.122	1.060	1.527	1.893
Bias	0.081	1.284	1.873	0.226	0.802	0.7	0.937	0.893
MSE	4.807	6.194	5.299	4.332	5.916	5.818	4.410	3.909
Length of asymptotic CI	4.986	5.722	6.413	2.714	3.101	3.672	1.951	5.025
$\boldsymbol{n = 100}$	$\widehat{\boldsymbol{\alpha}}_{\mathbf{1}}$	$\widehat{\boldsymbol{\alpha}}_{\mathbf{2}}$	$\widehat{\boldsymbol{\alpha}}_{\mathbf{3}}$	$\widehat{\boldsymbol{\delta}}_{\mathbf{1}}$	$\widehat{\boldsymbol{\delta}}_{\mathbf{2}}$	$\widehat{\boldsymbol{\theta}}_{\mathbf{1}}$	$\widehat{\boldsymbol{\theta}}_{\mathbf{2}}$	$\widehat{\boldsymbol{\lambda}}$
Mean	3.058	8.005	4.207	1.986	1.585	0.631	0.984	1.048
Bias	0.028	0.055	0.107	0.026	0.265	0.271	0.394	0.048
MSE	3.937	5.027	4.731	3.908	5.182	3.922	4.008	3.029
Length of asymptotic CI	4.306	5.291	4.285	2.831	2.399	2.068	0.886	3.638

5. Application

5.1. Diagnostic Probabilities Data Set Analysis

In this data, three behavioral states of attitudes or "diseases" of students known under the generic title of "newmath syndrome" are investigated and recorded using diagnostic probabilities. A sample of 15 students take part in this study, where diagnostic probabilities are assigned by clinicians for variables algebritis, bilateral paralexia and calculus deficiency.

The performance of the Dirichlet-gamma and the newly developed Kummer–Dirichlet gamma distributions are investigated here to see if these are suitable models for this data set, where the data has a correlation matrix given by

$$\begin{bmatrix} 1 & -0.332 & -0.581 \\ -0.332 & 1 & -0.574 \\ -0.581 & -0.574 & 1 \end{bmatrix}.$$

The initial parameter values needed for this performance test are obtained through a grid search using R software. The initial parameter values for the Dirichlet-gamma distribution are given as $(\alpha_1, \alpha_2, \alpha_3, \delta_1, \theta_1, \delta_2, \theta_2) = (9, 5, 5, 9, 2.2, 4, 9.4)$ and $(\alpha_1, \alpha_2, \alpha_3, \delta_1, \theta_1, \delta_2, \theta_2, \lambda) = (7, 6, 5, 6, 3, 3, 3, 0.5)$ as initial values for the Kummer–Dirichlet gamma distribution. Goodness-of-fit measures such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are used to illustrate the overall performance of the Kummer–Dirichlet gamma and Dirichlet-gamma distribution, where the model with the lowest values of AIC and BIC measures is considered to preferred.

The results of Table 2 and Figure 6 illustrate that the Kummer–Dirichlet gamma distribution serves an alternative model for compositional data sets. Reference [12] illustrated

that the Dirichlet-gamma is flexible in modeling compositional data sets; however, in this example, it is shown that the additional parameter λ adds flexibility, covering outliers where the Dirichlet-gamma distribution might not reach. The maximum likelihood value (*ll*), and the AIC and BIC measures also proves that the Kummer–Dirichlet gamma is a better alternative for this data set.

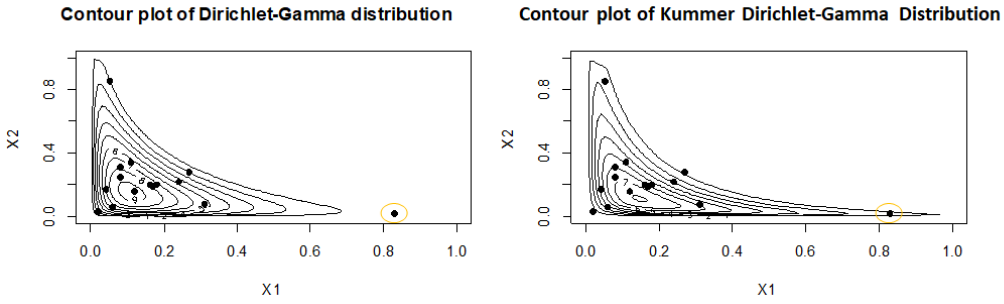

Figure 6. Contour plots of the Dirichlet-gamma and Kummer–Dirichlet gamma distributions on diagnostic probabilities data.

Table 2. Parameter estimates and the performance analysis for the diagnostic probabilities data set.

Model	**MLE**							
	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\lambda}$	$\widehat{\delta}_1$	$\widehat{\theta}_1$	$\widehat{\delta}_2$	$\widehat{\theta}_2$
DGa	14.804	1.931	2.734	n/a	0.153	0.949	0.836	0.268
KDGa	17.8259	11.806	5.028	0.524	0.178	0.109	0.216	0.025
	ll	AIC	BIC					
DGa	−22.174	58.348	63.305					
KDGa	−17.961	51.921	57.586					

5.2. The Mice Morris Water Maze Behavior Data Set Analysis

In this experiment, the time spent by rodents in the four different quadrants of a water maze is analyzed. The Morris water maze is a behavioral test mostly used on rodents (see [25]). The experiment begins by placing a rodent in a circular pool of water, where it is required to swim until it finds an escape platform in the pool. The aim of the experiment is to investigate the memory abilities and or memory loss of different rodents. Figure 7 illustrates the experiment. In this data, seven wild-type rodents are placed in a pool of water, where the time spent in the different quadrants is recorded.

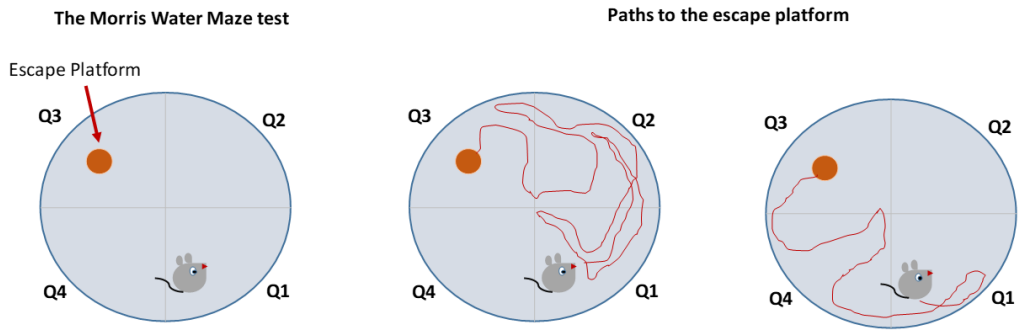

Figure 7. An illustration of the Morris water maze experiment.

In the study [25], the Dirichlet distribution was used as a suitable model for distinguishing the proportion of time spent across the different quadrants. In this example, the performance of Dirichlet distribution and the newly developed KDGa distribution is thus

compared to see if the KDGa distribution is superior, for this data set. The correlation matrix of this data is given by

$$\begin{bmatrix} 1 & -0.543 & -0.162 & -0.538 \\ -0.543 & 1 & -0.213 & -0.055 \\ -0.162 & -0.213 & 1 & -0.441 \\ -0.538 & -0.055 & -0.441 & 1 \end{bmatrix}. \tag{24}$$

The initial parameter values needed for this performance test, are obtained through a grid search using the R software. The initial parameter values for the Dirichlet distribution are given as $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (2, 2, 1, 4)$ and $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \delta_1, \theta_1, \delta_2, \theta_2, \delta_3, \theta_3, \lambda) = (1.27, 1.37, 1.20, 0.56, 1.15, 2.13, 0.84, 1.04, 1.06, 1.07, 1)$ as the initial values for the Kummer–Dirichlet gamma distribution.

Results of Table 3 illustrate that the Kummer–Dirichlet gamma distribution is a good competitor for this compositional data set. The estimation values of the parameters $(\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4)$ indicates the "weight" of each quadrant. For both the Dirichlet and the KDGa distribution, the value of $\widehat{\alpha}_1$ is higher than the values of $\widehat{\alpha}_2, \widehat{\alpha}_3, \widehat{\alpha}_4$, indicating that more time was spent in the first quadrant. The maximum likelihood value (ll), and the AIC and BIC measures also illustrate that the Kummer–Dirichlet gamma can be viewed as a good addition in analyzing this type of data set.

Table 3. Parameter estimates and performance analysis for the mice Morris water maze behavior data set.

Model	**MLE**								
	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\alpha}_3$	$\widehat{\alpha}_4$	$\widehat{\delta}_1$	$\widehat{\theta}_1$	$\widehat{\delta}_2$	$\widehat{\theta}_2$	$\widehat{\delta}_3$
Dirichlet	12.556	10.610	9.022	9.492	n/a	n/a	n/a	n/a	n/a
KDGa	1.493	1.378	1.202	0.563	1.147	2.130	0.844	1.045	1.061
	$\widehat{\theta}_3$	$\widehat{\lambda}$	ll	AIC	BIC				
Dirichlet	n/a	n/a	−29.513	65.026	64.864				
KDGa	1.075	1.000	−16.056	54.112	53.517				

6. Conclusions and Discussion

In this paper, the Kummer–Dirichlet gamma (KDGa) distribution is presented, which is a member of the proposed Kummer–Dirichlet (KD) class of distributions. It is illustrated that other distributions and their marginal distributions emanate from this class of distributions, of which include the Dirichlet-generated, with marginal beta-generated distributions and the exponentiated-generalized distribution as well. The pdf and moments of the KDGa distribution can be expressed as an infinite sum of that of the Dirichlet-gamma (DGa) distributions. The impact and usefulness of the KDGa distribution are illustrated via synthetic and real data sets, where its performance is compared to that of the Dirichlet and DGa distributions. We illustrated how this innovation of the Dirichlet distribution proposes a better fit for compositional psychology diagnostic data sets where outliers are present. The extra parameter λ of the KDGa distribution proves to add more flexibility in modeling compositional data sets.

To conclude this section, we will briefly discuss two other applications of the proposed KD generator model. A generative discriminative classifier can be well-defined by solving the following compound of KD with a multinomial distribution integral

$$\mathcal{KDCM}(\boldsymbol{X}|\boldsymbol{\alpha}) = \int_{\boldsymbol{y}} \mathcal{M}(\boldsymbol{X}|\boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y}$$

where $f(\cdot)$ is given by (8) and $\mathcal{M}(\boldsymbol{X}|\boldsymbol{y})$ is the pdf of a multinomial distribution with object $\boldsymbol{X}$ and parameters $\boldsymbol{y}$. See [26] for a recent similar approach. Another well-known application is in Bayesian analysis, where one can use the KD generator distribution as the prior in the multinomial distribution or allocation probabilities for clustering in a finite mixture model or either probabilistic graphical network modeling.

Author Contributions: Conceptualization, A.B. and M.A.; methodology, A.B., M.A. and S.M.; validation, S.M., M.A. and A.B.; formal analysis, S.M., A.B. and M.A.; investigation, S.M.; resources, S.M.; writing—original draft preparation, S.M.; writing—review and editing, M.A. and A.B.; visualization, S.M.; supervision, M.A. and A.B.; project administration, S.M.; funding acquisition, M.A. and A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was based upon research supported in part by the Visiting professor programme, University of Pretoria and the National Research Foundation (NRF) of South Africa, SARChI Research Chair UID: 71199; Reference: IFR170227223754 grant No. 109214; and Reference: SRUG190308422768 grant No. 120839. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF. The research of the corresponding author is supported by a grant from Ferdowsi University of Mashhad (N.2/55271).

Data Availability Statement: The data used in this article may be simulated in R, using the stated seed value and parameter values. The first real data set is available from the "compositional" package in R software, and the second real data set is available from the article referenced in [25].

Acknowledgments: We would like to sincerely thank two anonymous reviewers for their constructive comments, which led us to put many details in the paper and improved the presentation.

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

Appendix A. Proof of the Main Results

Proof of Theorem 1. Expanding the exponential term in (8), using the Taylor series, it follows that

$$\begin{aligned}
\frac{1}{C_2(\boldsymbol{\alpha},\lambda)} &= \int\limits_{\boldsymbol{\Omega}_p}\cdots\int\prod_{i=1}^{p} y_i^{\alpha_i-1}\left(1-\sum_{i=1}^{p} y_i\right)^{\alpha_{p+1}-1} e^{-\lambda\left(\sum_{i=1}^{p} y_i\right)}\,\boldsymbol{dy}\\
&= \sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)\int\limits_{\boldsymbol{\Omega}_p}\cdots\int\prod_{i=1}^{p} y_i^{\alpha_i+m_i-1}\left(1-\sum_{i=1}^{p} y_i\right)^{\alpha_{p+1}-1}\boldsymbol{dy}\\
&= \sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)\int\limits_{\boldsymbol{\Omega}_p}\cdots\int y_1^{\alpha_1+m_1-1}\prod_{i=2}^{p} y_i^{\alpha_i+m_i-1}\\
&\quad\times\left(1-y_1-\sum_{i=2}^{p} y_i\right)^{\alpha_{p+1}-1}\boldsymbol{dy}\\
&= \sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)\int\limits_{\boldsymbol{\Omega}_p}\cdots\int\left(\frac{y_1}{1-\sum_{i=2}^{p} y_i}\right)^{\alpha_1+m_1-1}\\
&\quad\times\left(1-\frac{y_1}{1-\sum_{i=2}^{p} y_i}\right)^{\alpha_{p+1}-1}\prod_{i=2}^{p} y_i^{\alpha_i+m_i-1}\left(1-\sum_{i=2}^{p} y_i\right)^{\alpha_1+m_1+\alpha_{p+1}-2}\boldsymbol{dy} \qquad \text{(A1)}
\end{aligned}$$

Let $q_1=\frac{y_1}{1-\sum_{i=2}^{p} y_i}$ in (A1), where $dy_1=\left(1-\sum_{i=2}^{p} y_i\right)dq_1$. Then

$$\begin{aligned}
\frac{1}{C_2(\boldsymbol{\alpha},\lambda)} &= \sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)B(\alpha_1+m_1,\alpha_{p+1})\int\limits_{\boldsymbol{\Omega}_{p-1}}\cdots\int\prod_{i=2}^{p} y_i^{\alpha_i+m_i-1}\\
&\quad\times\left(1-\sum_{i=2}^{p} y_i\right)^{\alpha_1+m_1+\alpha_{p+1}-2} dy_2\,dy_3\cdots dyp, \qquad \text{(A2)}
\end{aligned}$$

where $B(\cdot,\cdot)$ is the beta function. The expression in (A2) can further be simplified by a continuous process of a change of variable $q_j = \frac{y_j}{1-\sum_{j=i+1}^{p} y_j}$ for $j = 1, 2, \cdots, p$. The result of (A2) is solved in detail by [12]. Hence,

$$\begin{aligned}\frac{1}{C_2(\boldsymbol{\alpha},\lambda)} &= \sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)\prod_{i=1}^{p-1} B\left(\alpha_i + m_i,\ \sum_{j=i+1}^{p}\alpha_j + m_j + \alpha_{p+1}\right)\\ &\quad \times B\left(\alpha_p + m_p, \alpha_{p+1}\right). \end{aligned} \tag{A3}$$

It then follows that

$$\begin{aligned}\frac{1}{C_2(\boldsymbol{\alpha},\beta)} &= \sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)\prod_{i=1}^{p-1}\frac{\Gamma(\alpha_i+m_i)\Gamma(\sum_{j=i+1}^{p}\alpha_j + m_j + \alpha_{p+1})}{\Gamma(\alpha_i + m_i + \sum_{j=i+1}^{p}\alpha_j + m_j + \alpha_{p+1})}\\ &\quad \times\frac{\Gamma(\alpha_p+m_p)\Gamma(\alpha_{p+1})}{\Gamma(\alpha_p+\alpha_{p+1}+m_p)}\\ &= \sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)\prod_{i=1}^{p-1}\frac{\Gamma(\alpha_i)\ \Gamma(\sum_{j=i+1}^{p}\alpha_j+\alpha_{p+1})}{\Gamma(\alpha_i+\sum_{j=i+1}^{p}\alpha_j+\alpha_{p+1})}\times\frac{\Gamma(\alpha_p)\Gamma(\alpha_{p+1})}{\Gamma(\alpha_p+\alpha_{p+1})}\\ &\quad\times\left(\prod_{i=1}^{p-1}\frac{(\alpha_i)_{m_i}\ (\sum_{j=i+1}^{p}\alpha_j+\alpha_{p+1})_{\sum_{j=i+1}^{p} m_j}}{(\alpha_i+\sum_{j=i+1}^{p}\alpha_j+\alpha_{p+1})_{m_i+\sum_{j=i+1}^{p} m_j}}\right)\frac{(\alpha_p)_{m_p}}{(\alpha_p+\alpha_{p+1})_{m_p}}\\ &= \frac{\prod_{i=1}^{p}\Gamma(\alpha_i)\Gamma(\alpha_{p+1})}{\Gamma(\sum_{i=1}^{p}\alpha_i+\Gamma(\alpha_{p+1}))}\sum_{m_1,\cdots,m_p\geq 0}^{\infty}\left(\frac{-\lambda^{\sum_{i=1}^{p} m_i}}{\prod_{i=1}^{p} m_i!}\right)\frac{\prod_{i=1}^{p}(\alpha_i)_{m_i}}{(\sum_{i=1}^{p}\alpha_i+\alpha_{p+1})_{\sum_{i=1}^{p} m_i}}, \end{aligned}$$

which completes the proof, where α_i for $i = 1, 2, \cdots, p$ and where the summation of the Pochhammer functions can be represented as the confluent hypergeometric function ${}_1F_1(.;.;.)$. □

Proof of Theorem 2. By definition and using (10) and (13), it follows that

$$\begin{aligned} & M_{\boldsymbol{X}}(\boldsymbol{t}) = E\left[e^{\boldsymbol{t}\boldsymbol{X}^{\top}}\right]\\ &= C_2(\boldsymbol{\alpha},\lambda)\sum_{m_j,v_j,k\geq 0}^{\infty} w_{m_j,v_j,k}\int_{\mathcal{R}^p}\cdots\int\prod_{i=1}^{p} e^{t_i x_i}\, g_i(x_i)\, G_i^{\alpha_i+m_i+v_i-1}(x_i)\, \boldsymbol{dx}\\ &= C_2(\boldsymbol{\alpha},\lambda)\sum_{m_j,v_j,k\geq 0}^{\infty} w_{m_j,v_j,k}\int_{\mathcal{R}^p}\cdots\int\prod_{i=1}^{p}\frac{1}{\theta_i^{\delta_i}\Gamma(\delta_i)} e^{-x_i(\frac{1}{\theta_i}-t_i)}\ G_i^{\alpha_i+m_i+v_i-1}(x_i)\, \boldsymbol{dx}\\ &= C_2(\boldsymbol{\alpha},\lambda)\sum_{m_j,v_j,k\geq 0}^{\infty} w_{m_j,v_j,k}\int_{\mathcal{R}^p}\cdots\int\prod_{i=1}^{p}\frac{(\frac{1}{\theta_i}-t_i)^{\delta_i}}{\theta_i^{\delta_i}(\frac{1}{\theta_i}-t_i)^{\delta_i}\Gamma(\delta_i)} e^{-x_i(\frac{1}{\theta_i}-t_i)}\\ &\quad\times\ G_i^{\alpha_i+m_i+v_i-1}(x_i)\, \boldsymbol{dx}\\ &= C_2(\boldsymbol{\alpha},\lambda)\sum_{m_j,v_j,k\geq 0}^{\infty} w_{m_j,v_j,k}\times E\left\{\prod_{i=1}^{p}\frac{1}{\theta_i^{\delta_i}(\frac{1}{\theta_i}-t_i)^{\delta_i}} G_i^{\alpha_i+m_i+v_i-1}(Y_i)\right\}, \end{aligned} \tag{A4}$$

where $\boldsymbol{dx} = (dx_1 dx_2 ... dx_p)$ and $Y_i \sim Gamma(\frac{\theta_i}{1-\theta_i t_i}, \delta_i)$. Since $G_i(Y_i) \in (0,1)$, then let $G_i(Y_i) \equiv U_i \sim Unif(0,1)$. The proof is completed by simplifying the expected value in (A4) as follows

$$\begin{aligned} E\left\{\prod_{i=1}^{p}\frac{1}{\theta_i^{\delta_i}(\frac{1}{\theta_i}-t_i)^{\delta_i}} G_i^{\alpha_i+m_i+v_i-1}(Y_i)\right\} &= \prod_{i=1}^{p}\frac{1}{\theta_i^{\delta_i}(\frac{1}{\theta_i}-t_i)^{\delta_i}}\int_0^1\cdots\int_0^1 U_i^{\alpha_i+m_i+v_i-1}\, \boldsymbol{dU}\\ &= \prod_{i=1}^{p}\frac{1}{\theta_i^{\delta_i}(\frac{1}{\theta_i}-t_i)^{\delta_i}}\frac{1}{\alpha_i+m_i+v_i}, \end{aligned} \tag{A5}$$

where $\boldsymbol{dU} = (dU_1 dU_2 ... dU_p)$. The result of (18) follows from (A5). □

Proof of Theorem 3. For random vector $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathcal{R}^p$ with pdf (10) (represented as (13)), for $\sum_{i=1}^{p} G_i(x_i) < 1$, it follows that

$$\begin{aligned}
\mathcal{E} &= \int \cdots \int_{\mathcal{R}^p} C_2(\boldsymbol{\alpha}, \lambda) \left(1 - \sum_{i=1}^{p} G_i(x_i)\right)^{\alpha_{p+1}-1} \prod_{i=1}^{p} x_i^{n_i} g_i(x_i)\, G_i^{\alpha_i - 1}(x_i)\, e^{-\lambda G_i(x_i)}\, \boldsymbol{dx} \\
&= C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_j, v_j, k \geq 0}^{\infty} w_{m_j, v_j, k} \int \cdots \int_{\mathcal{R}^p} \prod_{i=1}^{p} x_i^{n_i}\, g_i(x_i)\, G_i^{\alpha_i + m_i + v_i - 1}(x_i)\, \boldsymbol{dx} \\
&= C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_j, v_j, k \geq 0}^{\infty} w_{m_j, v_j, k} \int \cdots \int_{\mathcal{R}^p} \prod_{i=1}^{p} \frac{1}{\theta_i^{\delta_i}\Gamma(\delta_i)} x_i^{n_i + \delta_i} e^{-\frac{x_i}{\theta_i}}\, G_i^{\alpha_i + m_i + v_i - 1}(x_i)\, \boldsymbol{dx} \\
&= C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_j, v_j, k \geq 0}^{\infty} w_{m_j, v_j, k} \int \cdots \int_{\mathcal{R}^p} \prod_{i=1}^{p} \frac{\theta_i^{n_i}\Gamma(n_i + \delta_i)}{\Gamma(\delta_i)} \frac{1}{\theta_i^{n_i + \delta_i}\Gamma(n_i + \delta_i)} \\
&\quad \times \; x_i^{n_i + \delta_i} e^{-\frac{x_i}{\theta_i}}\, G_i^{\alpha_i + m_i + v_i - 1}(x_i)\, \boldsymbol{dx} \\
&= C_2(\boldsymbol{\alpha}, \lambda) \sum_{m_j, v_j, k \geq 0}^{\infty} w_{m_j, v_j, k} \times E\left\{ \prod_{i=1}^{p} \frac{\theta_i^{n_i}\Gamma(n_i + \delta_i)}{\Gamma(\delta_i)} G_i^{\alpha_i + m_i + v_i - 1}(Y_i) \right\},
\end{aligned} \tag{A6}$$

where $\boldsymbol{dx} = (dx_1 dx_2 ... dx_p)$, $Y_i \sim Gamma(\theta_i, n_i + \delta_i)$ and $G_i(Y_i) \equiv U_i \sim Unif(0,1)$. The proof is completed by simplyifing the expected value in (A6), following the same procedure as in (A5). □

References

1. Aitchison, J. The statistical analysis of compositional data (with discussion). *J. R. Stat. Soc. Ser. B* **1982**, *44*, 139–177.
2. Balakrishnan, N.; Nevzorov, V.B. *A Primer on Statistical Distributions*; John Wiley & Sons: New York, NY, USA, 2003.
3. Barndorff-Nielsen, O.E.; Jorgensen, B. Some parametric models on the simplex. *J. Multivar. Anal.* **1991**, *39*, 106–116. [CrossRef]
4. Connor, J.R.; Mosimann, J.E. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *J. Am. Stat. Assoc.* **1969**, *64*, 194–206. [CrossRef]
5. Epaillard, A.; Bouguila, N. Data-free metrics for Dirichlet and generalized Dirichlet mixture-based HMMs-A practical study. *Pattern Recognit.* **2019**, *8*, 207–219. [CrossRef]
6. Favaro, S.; Hadjicharalambous, G.; Prunster, I. On a class of distributions on the simplex. *J. Stat. Plan. Inference* **2011**, *141*, 2987–3004. [CrossRef]
7. Ng, K.W.; Tian, G.L.; Tang, M.L. *Dirichlet and Related Distributions; Theory, Methods and Applications*; John Wiley & Sons: New York, NY, USA, 2011; Volume 1.
8. Thomas, S.; Jacob, J. A Generalized Dirichlet model. *Stat. Probab. Lett.* **2006**, *76*, 1761–1767. [CrossRef]
9. Marshall, A.; Olkin, I. *Inequalities: Theory of Majorization and Its Applications*; Academic Press: New York, NY, USA, 1979.
10. Gupta, R.D. Generalized Liouville Distributions. *Comput. Math. Appl.* **1996**, *32*, 103–109. [CrossRef]
11. Sivazlian, B.D. On a Multivariate extension of the Gamma and Beta distributions. *J. Appl. Math. Soc. Ind. Appl. Math.* **1981**, *41*, 205–209. [CrossRef]
12. Arashi, M.; Bekker, A.; de Waal, D.J.; Makgai, S.L. Constructing multivariate distributions via the Dirichlet generator. In *Computational and Methodological Statistics and Biostatistics. Contemporary Essays in Advancement*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 159–186.
13. Eugene, N.; Lee, C.; Famoye, F. Beta-normal distribution and its applications. *Commun. Stat.-Theory Methods* **2002**, *31*, 497–512. [CrossRef]
14. Ng, K.W.; Kotz, S. *Kummer-Gamma and Kummer-Beta Univariate and Multivariate Distributions*; Research Report, 84; Department of Statistics, The University of Hong Kong: Hong Kong, China, 1995.
15. Pescim, R.R.; Cordeiro, G.M.; Demetrio, C.G.B.; Ortega, E.M.M.; Nadarajah, S. The new class of Kummer beta generalized distributions. *Stat. Oper. Trans.* **2012**, *36*, 153–180.
16. Pescim, R.R.; Cordeiro, G.M.; Nararajah, S.; Demetrio, C.G.B.; Ortega, E.M.M. The Kummer beta Birnbaum-Saunders: An alternative fatigue life distribution. *Hacet. J. Math. Stat.* **2014**, *43*, 473–510.
17. Bran-Cardona, P.A.B.; Orozco-Castaneda, J.M.; Nagar, D.K. Bivariate Generalization of the Kummer-Beta Distribution. *Revista Colombiana de Estadistica* **1969**, *34*, 497–512.
18. Pescim, R.R.; Nararajah, S. The Kummer Beta Normal: A New Useful-Skew Model. *J. Data Sci.* **2015**, *13*, 509–532. [CrossRef]

19. Cordeiro, G.M.; Pescim, R.R.; Demetrio, C.G.B.; Ortega, E.M.M. The Kummer Beta Generalized Gamma Distribution. *J. Data Sci.* **2014**, *12*, 661–698. [CrossRef]
20. Mudholkar, G.S.; Srivastava, D.K.; Friemer, M. The exponential Weibull family: A reanalysis of the bus-motor failure data. *Technometrics* **1995**, *37*, 436–445. [CrossRef]
21. Gupta, R.C.; Gupta, P.L.; Gupta, R.D. Modeling failure time data by Lehmann alternatives. *Commun. Stat.-Theory Methods* **1998**, *27*, 887–904. [CrossRef]
22. Gupta, A.K.; Kundu, D. Exponentiated exponential family: An alternative to gamma and Weibull distributions. *Biom. J.* **2001**, *43*, 117–130. [CrossRef]
23. Mudholkar, G.S.; Srivastava, D.K.; Friemer, M. Exponentiated Weibull family for analyzing bathtub failure real data. *IEEE Trans. Reliab.* **1993**, *42*, 299–302. [CrossRef]
24. Bain, L.J.; Engelhardt, M. *Introduction to Probability and Mathematical Statistics*, 2nd ed.; Brooks/Cole Cengage Learning: Boston, MA, USA, 1992.
25. Maugard, M.; Doux, C.; Bonvento, G.A. new statistical method to analyze Morris Water Maze data using Dirichlet distribution. *F1000Research* **2019**, *8*, 1–14. [CrossRef] [PubMed]
26. Zamzamy, N.; Bouguila, N. Hybrid generative discriminative approaches based on multinomial scaled Dircihlet. *Appl. Intell.* **2019**, *49*, 3783–3800. [CrossRef]