Positive Orthant Hyperspherical Distribution and Applications

Jose H. Guardiola*

Abstract

In text mining, gene expressions and machine learning there is a need to model vectors at the positive orthant of the hypersphere. Similarly, in compositional data analysis a square root transformation can also be used to map the simplex onto the mentioned subspace. This paper focuses in developing a probability distribution on that region avoiding unnecessary probability mass at the whole hypersphere. We modified a proposed spherical Dirichlet distribution proposing a flexible version of this distribution. The distribution basic properties, such as normalizing constants and moments are developed. Efficient estimators based on classical inferential statistics are also obtained. An application using simulated data and a text mining example are developed and their results are discussed.

Key Words: Dirichlet distribution, text mining, hypersphere, gene expressions, positive orthant

1. Introduction

In text mining and gene expressions analysis, the collections of texts are represented in a vector-space model, which implies that texts once standardized, are coded as vectors in a sphere of higher dimensions, also called a hypersphere [9]. Many researchers currently model those distributions by means of existing probability density mixtures, however, these approximations waste probability mass in the whole hypersphere, when it is actually only needed at the positive orthant of the hypersphere. This is mainly because of the non-existence of suitable distributions for that subspace. The new proposed distribution fills that void, allowing a more efficient modeling of these vectors.

1.1 Probability Density Function and Normalizing Constants

The spherical-Dirichlet distribution is obtained by transforming the Dirichlet distribution on the simplex to the corresponding space on the hypersphere. In this section we derive the density and we compute the normalizing constants. Let $\boldsymbol{y}$ have a Dirichlet distribution on the simplex as described by Ingram [8].

$$\begin{aligned} f_{\text{Dir}}(\mathbf{y};\alpha) &= \frac{\Gamma(\sum_{i=1}^{m}\alpha_i)}{\prod_{i=1}^{m}\Gamma(\alpha_i)}\prod_{i=1}^{m} {y_i}^{\alpha_i-1}. \\ &= \frac{\Gamma(\sum_{i=1}^{m}\alpha_i)}{\prod_{i=1}^{m}\Gamma(\alpha_i)}\prod_{i=1}^{m-1} {y_i}^{\alpha_i-1}(1-\sum_{i=1}^{m-1} y_i)^{(\alpha_m-1)} \end{aligned} \tag{1}$$

where

$$0 \leqq y_i \leqq 1, \;\; \sum_{i=1}^{m} y_i = 1, \;\; \alpha_i \in \Re$$

*Texas A&M University-Corpus Christi, Department of Mathematics and Statistics

Transforming the Dirichlet distribution (1) from the simplex to the positive orthant of the hypersphere (Figure 1)

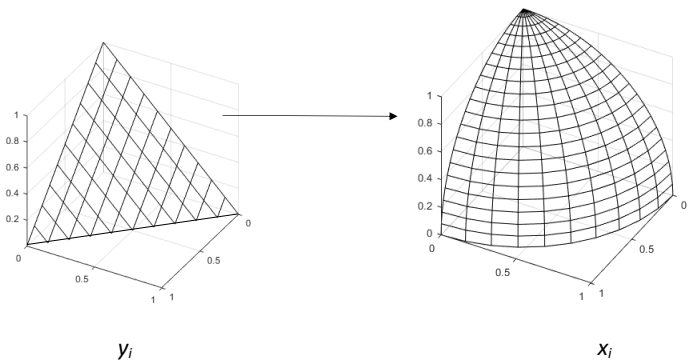

Figure 1: Transform from the simplex to the positive orthant of the hypersphere

and taking the square root transformation

$$x_i = \sqrt{y_i}, \;\; y_i = {x_i}^2, \;\; \frac{\partial y_i}{\partial x_i} = 2x_i \;\; i = 1, ...m \tag{2}$$

then, computing the Jacobian for all variables, it follows that

$$J = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} = 2x_1 & \frac{\partial y_1}{\partial x_2} = 0 & 0 & \dots \\ \frac{\partial y_2}{\partial x_1} = 0 & \frac{\partial y_2}{\partial x_2} = 2x_2 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{\partial y_m}{\partial x_m} = 2x_m \end{vmatrix} = \prod_{i=1}^{m} 2x_i = 2^m \prod_{i=1}^{m} x_i$$

the proposed transformation results in

$$\begin{aligned} f_{\text{SDir}}(\mathbf{x}; \alpha) &= \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m} {x_i}^{2\alpha_i - 1} \\ &= \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^{m} \Gamma(\alpha_i)} \prod_{i=1}^{m-1} {x_i}^{2\alpha_i - 1} (1 - \sum_{i=1}^{m-1} x_i^2)^{(\alpha_m - \frac{1}{2})} \end{aligned} \tag{3}$$

where

$$\sum_{i=1}^{m} \alpha_i =: \alpha_0, \;\; 0 \leqq x_i \leqq 1, \;\; \sum_{i=1}^{m} x_i^2 = 1, \;\; \alpha_i \in \Re^+.$$

We refer to (3) as the spherical Dirichlet distribution (SDD) and write $x \sim SDD(\alpha_i)$. We introduce the parameters α_i as the concentration parameters in a similar manner to the corresponding parameters of the Dirichlet distribution.

1.2 Moments

In this section we compute the first and second order moments, mode, standard deviation, variances and covariances and its corresponding covariance matrix. First, we compute the expected value for one of the variables, for example let us consider the expected value of x_1

$$E(x_1) = \int \cdots \int \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} x_1 (\prod_{i=1}^m {x_i}^{2\alpha_i - 1}) \mathrm{d}x_1 \dots \mathrm{d}x_m \tag{4}$$

$$= \int \cdots \int \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} x_1 (\prod_{i=1}^{m-1} {x_i}^{2\alpha_i - 1}) (1 - \sum_{i=1}^{m-1} x_i^2)^{(\alpha_m - \frac{1}{2})} \mathrm{d}x_1 \dots \mathrm{d}x_{m-1}, \tag{5}$$

where we recognize the expression inside the integral as the kernel of the proposed SDD with a new first parameter $\alpha_1 + \frac{1}{2}$ replacing the original α_1, then we can rewrite immediately this expression as

$$E(x_1) = \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} \frac{\Gamma(\alpha_1 + \frac{1}{2}) \prod_{i=2}^m \Gamma(\alpha_i)}{2^m \Gamma(\alpha_0 + \frac{1}{2})} \tag{6}$$

or equivalently,

$$E(x_1) = \frac{\mu_1}{\mu_0}, \tag{7}$$

where we write μ_i as,

$$\mu_i =: \frac{\Gamma(\alpha_i + \frac{1}{2})}{\Gamma(\alpha_i)}. \tag{8}$$

The general solution for the first moment for a vector $\mathbf{x}$ can be written as

$$E(\mathbf{x}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + \frac{1}{2})} \left(\frac{\Gamma(\alpha_1 + \frac{1}{2})}{\Gamma(\alpha_1)}, \dots, \frac{\Gamma(\alpha_m + \frac{1}{2})}{\Gamma(\alpha_m)} \right) = \frac{1}{\mu_0} \frac{\Gamma(\boldsymbol{\alpha} + \frac{1}{2})}{\Gamma(\boldsymbol{\alpha})} \tag{9}$$

that can also be written

$$\boldsymbol{\mu} = \frac{\Gamma(\boldsymbol{\alpha} + \frac{1}{2})}{\Gamma(\boldsymbol{\alpha})}, \tag{10}$$

then the expected value for a vector $\boldsymbol{x}$ is

$$E(\mathbf{x}) = \frac{\boldsymbol{\mu}}{\mu_0} = \frac{||\boldsymbol{\mu}||}{\mu_0} \cdot \frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||} = C \cdot \bar{\boldsymbol{\mu}}, \tag{11}$$

let

$$C =: \frac{||\boldsymbol{\mu}||}{\mu_0}, \quad \bar{\boldsymbol{\mu}} =: \frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}, \quad \bar{\boldsymbol{\mu}} \in \Omega_{m-1}. \tag{12}$$

Similarly, computing the expected value for x_1^2

$$E(x_1^2) = \int \cdots \int \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} x_1^2 \cdot (\prod_{i=1}^m {x_i}^{2\alpha_i - 1}) \mathrm{d}x_1 \ldots \mathrm{d}x_m, \tag{13}$$

then

$$E(x_1^2) = \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} \int \cdots \int x_1^{2\alpha_1 - 1} (\prod_{i=2}^m {x_i}^{2\alpha_i - 1}) \mathrm{d}x_1 \ldots \mathrm{d}x_{m-1}, \tag{14}$$

again, we can recognize the expression inside the integral as the kernel of the proposed SDD with a new first parameter $\alpha_1 + 1$, that yields

$$E(x_1^2) = \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} \frac{\Gamma(\alpha_1 + 1) \prod_{i=2}^m \Gamma(\alpha_i)}{2^m \Gamma(\alpha_0 + 1)} = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + 1)} \frac{\Gamma(\alpha_1 + 1)}{\Gamma(\alpha_1)} = \frac{\alpha_1}{\alpha_0}, \tag{15}$$

this result can be generalized to any x_i as

$$E(x_i^2) = \frac{\alpha_i}{\alpha_0}. \tag{16}$$

Moreover, the variance for any variable x_i is

$$V(x_i) = \frac{\alpha_i}{\alpha_0} - \frac{\mu_i^2}{\mu_0^2}, \tag{17}$$

and the covariance for x_1, x_2 can be written as

$$E(x_1 \cdot x_2) = \int \cdots \int \frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} x_1 \cdot x_2 (\prod_{i=1}^m {x_i}^{2\alpha_i - 1}) \mathrm{d}x_1 \ldots \mathrm{d}x_m, \tag{18}$$

after some arrangements, we can identify the kernel of the proposed SDD with the first two parameters as $\alpha_1 + \frac{1}{2}$, and $\alpha_2 + \frac{1}{2}$, where we can solve the corresponding integral, and our result takes the form

$$E(x_1 \cdot x_2) = \frac{\Gamma(\alpha_1 + \frac{1}{2}) \Gamma(\alpha_2 + \frac{1}{2})}{\alpha_0 \Gamma(\alpha_1) \Gamma(\alpha_2)} = \frac{\mu_1 \cdot \mu_2}{\alpha_0}. \tag{19}$$

In general for any pair of variables (x_i, x_j) we can write

$$E(x_i \cdot x_j) = \delta_{ij} \cdot \frac{\alpha_i}{\alpha_0} + (1 - \delta_{ij}) \cdot \frac{\mu_i \cdot \mu_j}{\alpha_0}, \tag{20}$$

where δ_{ij} is the delta Kronecker, and we can also write the covariance for any pair of variables (x_i, x_j) as

$$COV(x_i, x_j) = \left(\frac{1}{\alpha_0} - \frac{1}{\mu_0^2}\right) \mu_i \cdot \mu_j \text{ for } i \neq j. \tag{21}$$

In general we can write the covariance for any pair of variables (x_i, x_j) as

$$COV(x_i, x_j) = \delta_{ij} \cdot \left(\frac{\alpha_{i=j}}{\alpha_0} - \frac{\mu_i^2}{\mu_0^2}\right) + (1 - \delta_{ij}) \cdot \left(\frac{1}{\alpha_0} - \frac{1}{\mu_0^2}\right) \mu_i \cdot \mu_j, \tag{22}$$

that in matrix notation can also be written as

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{\alpha_1}{\alpha_0} - \frac{\mu_1^2}{\mu_0^2} & \left(\frac{1}{\alpha_0} - \frac{1}{\mu_0^2}\right) \mu_1 \cdot \mu_2 & \dots & \dots \\ \left(\frac{1}{\alpha_0} - \frac{1}{\mu_0^2}\right) \mu_2 \cdot \mu_1 & \frac{\alpha_2}{\alpha_0} - \frac{\mu_2^2}{\mu_0^2} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \frac{\alpha_m}{\alpha_0} - \frac{\mu_m^2}{\mu_0^2} \end{bmatrix},$$

an equivalent expression is

$$\boldsymbol{\Sigma} = \frac{1}{\alpha_0} \begin{bmatrix} \alpha_1 - \mu_1^2 & 0 & \dots & \dots \\ 0 & \alpha_2 - \mu_2^2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \alpha_m - \mu_m^2 \end{bmatrix} - \left(\frac{1}{\mu_0^2} - \frac{1}{\alpha_0}\right) \boldsymbol{\mu}\boldsymbol{\mu}^T,$$

similarly we let

$$\boldsymbol{\Sigma} = \frac{1}{\alpha_0} diag(\boldsymbol{\alpha}) - \frac{C^2 \mu_0^2}{\alpha_0} diag(\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T) - C^2 \left(1 - \frac{\mu_0^2}{\alpha_0}\right) \bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T, \tag{23}$$

where

$$C = \frac{||\boldsymbol{\mu}||}{\mu_0}, \quad \bar{\boldsymbol{\mu}} = \frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}, \quad \bar{\boldsymbol{\mu}} \in \Omega_{m-1}, \tag{24}$$

that summarizes our results in a succinct form.

1.3 Mode and Relationship with the Mean

The mode for the SDD can be determined by finding the values of α_i that maximize this function, alternatively, we can also maximize the log of this function as it is customary and usually easier. First, taking the natural log of the SDD and adding the constraint $\sum_{i=1}^m x_i^2 = 1$ for the purpose of using Lagrange multipliers we get

$$\ln f_{\text{SDir}}(\mathbf{x}, \boldsymbol{\alpha}) = \ln\left(\frac{2^m \Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)}\right) + \sum_{i=1}^m (2\alpha_i - 1) \ln x_i - \lambda\left(\sum_{i=1}^m x_i^2 - 1\right), \tag{25}$$

taking derivatives with respect to x_i and setting them to zero we have

$$\frac{\partial \ln f_{\text{SDir}}}{\partial x_i} = (2\alpha_i - 1)\frac{1}{x_i} - 2x_i\lambda = 0, \tag{26}$$

solving for x_i^2, it yields

$$x_i^2 = \frac{2\alpha_i - 1}{2\lambda}, \tag{27}$$

and substituting this result at the constraint in (25) we can solve for λ as

$$\lambda = \frac{1}{2}(2\sum_{i=1}^{m}\alpha_i - m), \tag{28}$$

where we can obtain the mode for x_i as

$$x_i = \sqrt{\frac{2\alpha_i - 1}{2\alpha_0 - m}}. \tag{29}$$

Considering the special case of a symmetric SDD for all $\alpha_i = \alpha$, it yields for all x_i

$$\textit{(mode)}x_i = \sqrt{\frac{2\alpha - 1}{m \cdot (2\alpha - 1)}} = \frac{1}{\sqrt{m}} \text{ for } \forall\alpha_i = \alpha \text{ and } \alpha \neq \frac{1}{2}, \tag{30}$$

the mean for a symmetric Spherical-Dirichlet distribution for all $\alpha_i = \alpha$ is

$$E(x_i) = \frac{\mu_i}{\mu_0} = \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + \frac{1}{2})} = \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)} \cdot \frac{\Gamma(m\alpha)}{\Gamma(m\alpha + \frac{1}{2})}, \tag{31}$$

where we can see that the mode does not match the expected value for a symmetric SDD, however, we can still find an asympotic relationship using the expression developed by Frame [1]

$$\lim_{x\to\infty} f(x) = \frac{\Gamma(x + a)}{\Gamma(x)} = x^a, \tag{32}$$

using this approximation it yields

$$\lim_{\alpha\to\infty} E(x_i) = (\alpha^{\frac{1}{2}}) \cdot \frac{1}{(m\alpha)^{\frac{1}{2}}} = \frac{1}{\sqrt{m}} \text{ for } \forall\alpha_i = \alpha \text{ and } \alpha \neq \frac{1}{2}, \tag{33}$$

that in the limit it matches the mode shown in (30).

2. Relationships of the SDD with other Distributions

In this section we explore the relationships or the lack thereof, between the SDD and other popular distributions such as the uniform, von Mises and its particular case of the Fisher Bingham distribution. We consider limiting cases for different values of the concentration parameters α_i.

2.1 Limiting Case Symmetric Distribution for large α

Assuming a symmetric SDD with $\alpha_i = \alpha,\ \text{for}\ \forall\alpha_i$ we can write

$$f_{\text{Dir}}(\mathbf{y};\alpha) = \frac{\Gamma(m\alpha)}{\Gamma(\alpha)^m}\prod_{i=1}^{m} {x_i}^{2\alpha-1}, \tag{34}$$

subject to the restrictions

$$0 \leqq x_i \leqq 1,\ \ \sum_{i=1}^{m} x_i^2 = 1,\ \ \alpha \in \Re^+,$$

in this case the covariance matrix can be reduced to

$$\boldsymbol{\Sigma} = \frac{1}{m}\left(1 - \frac{\mu_\alpha^2}{\alpha}\right)\boldsymbol{I} - \left(\frac{\mu_\alpha}{\mu_0}\right)^2\left(1 - \frac{\mu_0^2}{m\alpha}\right)\mathbf{1}\mathbf{1}^T, \tag{35}$$

where

$$\mu_\alpha = \frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)}, \quad \mu_0 = \frac{\Gamma(\alpha_0 + \frac{1}{2})}{\Gamma(\alpha_0)},$$

in an attempt to write the SDD as a rotational distribution of the type shown by Mardia [5], the latter expression can be rewritten as

$$\boldsymbol{\Sigma} = \left(1 - \frac{\mu_\alpha^2}{\alpha}\right)\left(\frac{1}{m}\boldsymbol{I} - \bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T\right) + \left(1 - m\frac{\mu_\alpha^2}{\mu_0^2}\right)\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T, \tag{36}$$

or equivalently

$$\boldsymbol{\Sigma} = var(x)m\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T + \left(\frac{1 - \frac{\mu_\alpha^2}{\alpha}}{m}\right)\left(\boldsymbol{I} - m\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T\right), \tag{37}$$

where we can't determine an equivalence to the von-Mises or similar rotationally symmetric distributions, however, we can see that in the limiting case for $\alpha \to \infty$ and consequently $\alpha_0 \to \infty$ we have

$$\lim_{\alpha\to\infty}\mu_\alpha = \lim_{\alpha\to\infty}\frac{\Gamma(\alpha + \frac{1}{2})}{\Gamma(\alpha)} = \alpha^{\frac{1}{2}},$$

and

$$\lim_{\alpha\to\infty}\mu_0 = \lim_{\alpha\to\infty}\frac{\Gamma(m\alpha + \frac{1}{2})}{\Gamma(m\alpha)} = (m\alpha)^{\frac{1}{2}},$$

which in the limit it yields

$$\lim_{\alpha\to\infty}\boldsymbol{\Sigma} = \lim_{\alpha\to\infty}\left(1 - \frac{\mu_\alpha^2}{\alpha}\right)\left(\frac{1}{m}\boldsymbol{I} - \bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T\right) + \left(1 - m\frac{\mu_\alpha^2}{\mu_0^2}\right)\bar{\boldsymbol{\mu}}\bar{\boldsymbol{\mu}}^T = 0,$$

we can conclude that for large values of α the covariance matrix tends to zero, consequently, the SDD tends to be concentrated as a vector with no variation.

2.2 Limiting Case Uniform Distribution

We now consider the case where all $\alpha_i = \frac{1}{2}$, then the SDD becomes

$$f_{\text{SDir}}(\mathbf{x};\alpha) = \frac{2^m\Gamma(\alpha_0)}{\prod_{i=1}^m \Gamma(\alpha_i)} \prod_{i=1}^m {x_i}^{2\alpha_i-1} = \frac{2^m\Gamma(\frac{m}{2})}{\Gamma(\frac{1}{2})^m}, \tag{38}$$

which is a constant thickness independent of the values of x_i. Then the SDD becomes the uniform distribution over the positive orthant of the hypersphere.

2.3 Similarities and Differences of the SDD with the von Mises and Fisher Bingham Distributions

The von Mises distribution is usually considered the analogue of the normal distribution in the circle as described by Mardia in [4], and its particular case for the three dimensional sphere, the Fisher Bingham distribution, both tend to converge to a multivariate and bivariate normal distribution respectively for large values of κ as shown by Kent [2].

The proposed SDD doesn't seem to converge to the von Mises distribution or to a multivariate normal distribution for large values of α_i, but rather it tends to be concentrated as a vector as it was established at the end of Section 2.1.

Moreover, both the von Mises and the Fisher Bingham distribution converge to the uniform distribution for very small values of κ, in a similar way as the SDD does for all $\alpha_i = \frac{1}{2}$, as it was shown at section 2.2.

3. Inference for the Spherical Dirichlet Distribution

We now consider estimation of the parameters of the SDD. Our main interest is to develop suitable procedures to estimate the set of parameters α_i, given a sample of random vectors in the positive orthant of the hypersphere. We first derive estimators for α_i using the method of moments (MOM), next we develop estimators for the same set of parameters using the method of maximum likelihood estimation (MLE).

3.1 Method of Moments

Using a similar procedure as the one developed by Narayanan [7] to estimate the parameters of the Dirichlet distribution, suppose we have a random sample with *n* random vectors $X_1, X_2, ...X_n$ such that $X_i \in \Re^m = \left[X_j | j = 1..., m; X_j > 0, \sum_{j=1}^m x_j^2 = 1\right]$ that are i.i.d., then

$$E(x_i) = \frac{\Gamma(\alpha_i + \frac{1}{2})}{\Gamma(\alpha_i)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + \frac{1}{2})} = \frac{\mu_i}{\mu_0}, \tag{39}$$

and

$$E(x_i^2) = \frac{\alpha_i}{\alpha_0}. \tag{40}$$

We define the sample moments as

$$X_{1j}^{'} = \frac{1}{n}\sum_{i=1}^n x_{ij} \quad j = 1, .., m, \tag{41}$$

and

$$X_{2j}^{'} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 \quad j = 1, .., m. \tag{42}$$

We have *m* first order moment equations and *m* second order moment equations to solve for *m* unknowns α_i. To avoid linear dependency and for the sake of simplicity we choose one of the first order moments and *m-1* second order moment equations

$$\frac{\Gamma(\alpha_1 + \frac{1}{2})}{\Gamma(\alpha_1)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + \frac{1}{2})} = \frac{1}{n}\sum_{i=1}^{n} x_{i1} = X_{11}^{'}, \tag{43}$$

then, the remaining *m-1* second order moment equations are

$$\frac{\alpha_i}{\alpha_0} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = X_{2j}^{'} \quad j = 2, ..., m. \tag{44}$$

There is no closed form solution for α_i in solving simultaneously (43) and (44), so we must solve numerically to obtain the corresponding method of moments estimators for α_i. Results from MOM can be used as initial values for the MLE that usually exhibit better statistical properties.

3.2 Maximum Likelihood Estimation

Suppose that we have a random sample of vectors on the positive orthant of the hypersphere, $X_1, X_2, ...X_n$, where $X_i \in \Re^m$ from SDD with pdf (3). Then, the log-likelihood is

$$\begin{aligned} \ln L(\boldsymbol{\alpha}) &= \ln \prod_{i=1}^{n} \frac{2^m \Gamma\left(\sum_{j=1}^{m} \alpha_j\right)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} {x_{ij}}^{2\alpha_j - 1} \\ &= \ln \prod_{i=1}^{n} 2^m \Gamma(\sum_{j=1}^{m} \alpha_j) \prod_{j=1}^{m} \Gamma(\alpha_j)^{-1} {x_{ij}}^{2\alpha_j - 1}. \end{aligned} \tag{45}$$

The parameters for a SDD can be estimated maximizing the log-likelihood function of the data, in a similar procedure as the one used by Minka for the Dirichlet distribution described at [6]. We can group all the constant terms as K, and we can rewrite all the products and sums as

$$\begin{aligned} \ln L(\boldsymbol{\alpha}) &= K + n \ln \Gamma(\sum_{j=1}^{m} \alpha_j) - n \sum_{j=1}^{m} \ln \Gamma(\alpha_j) + \sum_{i=1}^{n} \sum_{j=1}^{m} (2\alpha_j - 1) \ln x_{ij}, \\ &= K + n \left(\ln \Gamma(\sum_{j=1}^{m} \alpha_j) - \sum_{j=1}^{m} \ln \Gamma(\alpha_j) + \sum_{j=1}^{m} (2\alpha_j - 1) \frac{1}{n} \sum_{i=1}^{n} \ln x_{ij} \right), \end{aligned}$$

where the function that needs to be optimized after removing unnecessary constants is

$$F(\boldsymbol{\alpha}) = \ln\Gamma(\sum_{j=1}^{m}\alpha_j) - \sum_{j=1}^{m}\ln\Gamma(\alpha_j) + \sum_{j=1}^{m}(2\alpha_j - 1)\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_{ij}\right). \quad (46)$$

The gradient of the objective function can be obtained by differentiating the log-likelihood $\ln F(\alpha)$ with respect to α_k as

$$\nabla(F)_k = \frac{\partial F}{\partial \alpha_k} = \Psi(\sum_{j=1}^{m}\alpha_j) - \Psi(\alpha_k) + 2\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_{ik}\right), \quad (47)$$

where $\Psi =: \frac{d\ln\Gamma(x)}{dx}$ is the digamma function. The optimization is subject to the constraints $\alpha_i \geqq 0$. Because the SDD is a member of the exponential family this is a convex function and the observed sufficient statistic is equal to the expected sufficient statistic, where the latter is

$$E\left(x_k\right) = \frac{1}{2}\Psi(\alpha_k) - \frac{1}{2}\Psi(\sum_{j=1}^{m}\alpha_j), \quad (48)$$

and the observed sufficient statistic is

$$\frac{1}{n}\sum_{i=1}^{n}\ln x_{ij}. \quad (49)$$

That leads to the following iterative procedure

$$\Psi(\alpha_k^{new}) = \Psi(\sum_{j=1}^{m}\alpha_j^{old}) + 2\left(\frac{1}{n}\sum_{i=1}^{n}\ln x_{ik}\right). \quad (50)$$

Although the proposed procedure does not guarantee in general reaching a global maximum, updating successively (50) to maximize the log-likelihood equation provides reasonable results and convergence is typically fast.

4. Applications to Data

Lets now consider estimation of the parameters of the SDD. We first developed an example using simulated data generated from the proposed SDD with known parameters, and assuming to be unknown for the purpose of this estimation. Next, a second example was developed using a text mining example, with data obtained from a publicly available data set. Both examples were solved using MOM and MLE, applying the proposed techniques described at sections 3.1 and 3.2, and results obtained from both methods were compared.

4.1 Simulation Example

Four different simulations were performed each with 1,000 randomly generated values from a SDD in a three-dimensional space, with known proposed values of the parameters α_1, α_2 and α_3. Inferences to estimate the values of these parameters, assumed to be unknown, were

performed using MOM and MLE procedures developed in the previous sections 3.1 and 3.2. Graphs for the SDD corresponding to the proposed four different sets of parameters are shown in figures 2 and 3.

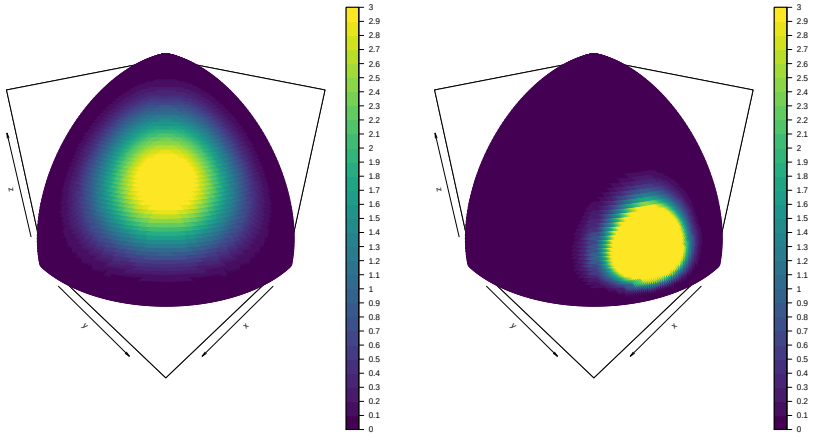

Figure 2: $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 2$ $\alpha_1 = 5, \alpha_2 = 15, \alpha_3 = 2$

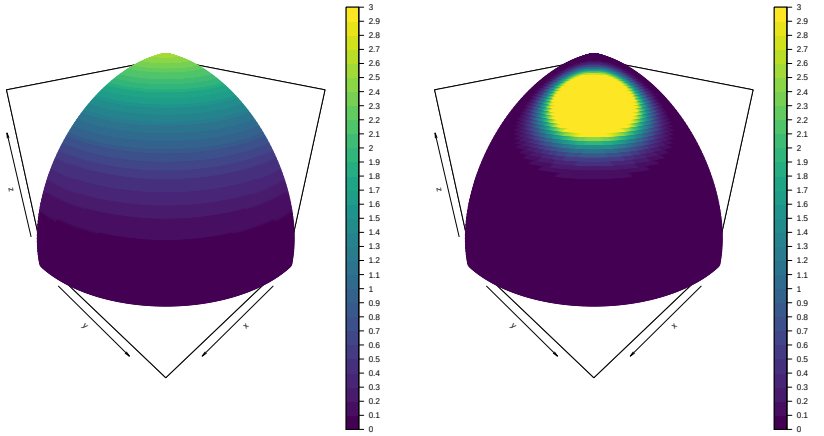

Figure 3: $\alpha_1 = 0.5, \alpha_2 = 0.5, \alpha_3 = 2$ $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 10$

First, an estimation was performed using MOM and iterating between (43) and (44). These values are updated in each cycle until convergence is achieved within the proposed tolerance limit. The estimated values of the parameters found using MOM were used as the initial values for the iterative process using MLE. For the latter method expression (50) is updated successively until the values of the parameters are stable within a pre-set tolerance level. Results for estimation by both methods and the true values of the parameters are shown at table 1. Note the close agreement between the MLEs and MOMs at the results shown at table 1.

Table 1: Simulation Results.

Method	**# Iterations**	$\boldsymbol{\alpha_1 = 2}$	$\boldsymbol{\alpha_2 = 2}$	$\boldsymbol{\alpha_3 = 2}$	**% Error**
MOM	176	2.0557	2.0983	2.0764	3.84
MLE	51	2.0412	2.0798	2.0479	2.81
Method	**# Iterations**	$\boldsymbol{\alpha_1 = 5}$	$\boldsymbol{\alpha_2 = 15}$	$\boldsymbol{\alpha_3 = 2}$	**% Error**
MOM	589	5.0351	14.7148	1.9684	1.40
MLE	147	5.1932	15.1998	2.0496	2.56
Method	**# Iterations**	$\boldsymbol{\alpha_1 = 0.5}$	$\boldsymbol{\alpha_2 = 0.5}$	$\boldsymbol{\alpha_3 = 2}$	**% Error**
MOM	28	0.4964	0.4639	1.9212	3.96
MLE	23	0.4903	0.4821	1.9503	2.67
Method	**# Iterations**	$\boldsymbol{\alpha_1 = 2}$	$\boldsymbol{\alpha_2 = 2}$	$\boldsymbol{\alpha_3 = 10}$	**% Error**
MOM	385	1.9349	2.0153	10.1145	1.72
MLE	85	1.9745	2.0713	10.3528	2.79

4.2 Text Mining Example

A text mining example was developed using a publicly available data set assembled by Lang [3]. An example of email messages regarding several interest groups are available, the "auto" topic was selected and summarized using standard data mining techniques. A collection of randomly selected 160 documents (emails) was extracted and summarized as vectors at the positive orthant of the hypersphere. Common terms such as "from" or "subject" were excluded as they did not provide any discriminant power and could potentially bias the analysis. Vocabulary reduction for synonymous and stemming were performed, and the ten most common terms were extracted by obtaining their raw frequencies. The frequencies for these terms can be expressed as vectors at a ten-dimensional space. A small sample of this data set can be seen at table 2.

Table 2: Terms Frequency.

Doc ID	**ntoken**	**auto**	**write**	**articl**	**engin**	**don**	**good**	**time**	**drive**	**road**
103092	0	2	1	1	0	0	0	0	0	0
101671	7	0	2	2	0	2	2	0	0	0
.....	...	...	...	...	...	...	...	...	...	...
101582	6	8	3	2	0	0	0	0	0	0
103050	0	3	1	0	0	0	0	0	0	0

An appropriate transformation for these vectors needs to be applied to reduce extreme values and eliminate zeros. The transformation that was applied here is $x_{transf} = \ln(1.10+ x)$. These vectors were standardized with a unit length at the positive quadrant of the hypersphere and they were fitted using the proposed multivariate SDD for ten dimensions. The estimation for the corresponding α's for the proposed distribution were done using both MOM and MLE, and their corresponding estimated values are shown at table 3.

Table 3: Text Mining Results.

Parameter	MOM	MLE
α_1	0.7799	1.1787
α_2	0.6545	1.0013
α_3	0.2151	0.4755
α_4	0.1790	0.4276
α_5	0.1182	0.2825
α_6	0.1268	0.3224
α_7	0.0923	0.2857
α_8	0.1054	0.3004
α_9	0.0833	0.2796
α_{10}	0.0591	0.2481

The number of iterations needed to fit the SDD within a preset tolerance level for MOM were 271. Using the MOM estimators as the initial values for MLE a new model was fitted using 19 additional iterations. Although the MLE procedure in general does not guarantee finding a global maximum, the proposed method provided reasonable results and the convergence was fast enough.

5. Conclusions

The proposed SDD is a good alternative to other methods for fitting unit vectors at the positive orthant of the hypersphere. The SDD avoids wasting probability mass or using distribution mixtures for the whole hypersphere. Results for MOM and MLE were in close agreement for simulated data, and reasonably close for a real text mining example. The simulated data was generated directly from the proposed SDD while the text mining data was obtained from a real practical problem. The SDD is flexible enough and it shows a rich variety of shapes that are able to fit a wide range of data, in a similar way as the beta distribution does for a one dimensional space. Future research will be aimed to enhance the capability of dealing with hyper-vectors that include zeros in some coordinates.

References

[1] J. S. Frame. An approximation to the quotient of gamma function. *The American Mathematical Monthly*, 56(8):529–535, 1949.

[2] John T. Kent. The Fisher-Bingham distribution on the sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982.

[3] Ken Lang. Cmu text learning group data archives. Available at https://www.cs.cmu.edu/afs/cs/project/theo-20/www/data/news20.html, 2019.

[4] K. V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(3):349–393, 1975.

[5] K.V. Mardia and P.E. Jupp. *Directional statistics*. Wiley series in probability and statistics. Wiley, 2000.

[6] Thomas P. Minka. Estimating a Dirichlet distribution. Technical Report, MIT. 2000.

[7] A. Narayanan. A note on parameter estimation in the multivariate beta distribution. *Computers and Mathematics with Applications*, 24(10):11–17, 1992.

[8] Ingram Olkin and Herman Rubin. Multivariate beta distributions and independence properties of the Wishart distribution. *The Annals of Mathematical Statistics*, 35(1):261–269, 3 1964.

[9] Sra Suvrit. Directional statistics in machine learning: a brief review. *arXiv e-prints*, page arXiv:1605.00316, 2016.