On The Dirichlet Distribution

by

Jiayu Lin

A report submitted to the
Department of Mathematics and Statistics
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
September 2016

Abstract

The Dirichlet distribution is a multivariate generalization of the Beta distribution. It is an important multivariate continuous distribution in probability and statistics. In this report, we review the Dirichlet distribution and study its properties, including statistical and information-theoretic quantities involving this distribution. Also, relationships between the Dirichlet distribution and other distributions are discussed. There are some different ways to think about generating random variables with a Dirchlet distribution. The stick-breaking approach and the Pólya urn method are discussed.

In Bayesian statistics, the Dirichlet distribution and the generalized Dirichlet distribution can both be a conjugate prior for the Multinomial distribution. The Dirichlet distribution has many applications in different fields. We focus on the unsupervised learning of a finite mixture model based on the Dirichlet distribution. The Initialization Algorithm and Dirichlet Mixture Estimation Algorithm are both reviewed for estimating the parameters of a Dirichlet mixture. Three experimental results are shown for the estimation of artificial histograms, summarization of image databases and human skin detection.

Acknowledgement

I would like to express my sincerest gratitude to my supervisors, Dr. Fady Alajaji and Dr. Glen Takahara. I appreciate their many thoughtful suggestions and support throughout the completion of this report. My thanks go to Queen's University and the Department of Mathematics and Statistics for offering me the opportunity to continue my studies in Statistics and related fields. Finally, I would like to thank my family and friends for their love, care and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Dirichlet distribution is an important multivariate continuous distribution in probability and statistics. As a multivariate generalization of the Beta distribution, the Dirichlet distribution is the most natural distribution for compositional data and measurements of proportions modeling [34]. In Bayesian statistics, the Dirichlet distribution is a popular conjugate prior for the Multinomial distribution.

There are many applications for the Dirichlet distribution in various fields. For example, the Dirichlet distirbution is used in deriving the distribution function of order statistics [40]. In biology, reference [28] demonstrates that the Dirichlet distribution can be used to compute forensic match probabilities from several distinct populations. Also, the Dirichlet distribution can be used to model a player's abilities in Major League Baseball [37]. In [23], it is shown that the Dirichlet distribution can be used to model consumer buying behaviour. An application that we focus on in this report (in Chapter 4) is the unsupervised learning of a finite mixture model based on

the Dirichlet distribution [12].

Extensions of the Dirichlet distribution are helpful to represent different purposes in various applications. For example, the Grouped Dirichlet distribution and the nested Dirichlet distribution can be used for statistical analysis of incomplete categorical data [34]. Also, there are some distributions related to the Dirichlet distribution, such as the generalized Dirichlet distribution, the hyper-Dirichlet distribution, the Dirichlet-Multinomial distribution, the scaled Dirichlet distribution and the mixed Dirichlet distribution [34].

1.1 Organization of Report

Chapter 2 reviews the definition of the Gamma distribution and the Beta distribution, which will be used in deriving the Dirichlet distribution. Also, the definition and some main properties of the Dirichlet distribution will be shown. More specifically, statistical and information-theoretic quantities involving the Dirichlet distribution are derived. Next, the stick-breaking approach is shown for generating Dirichlet distributed random vectors. Chapter 3 describes exchangeability and the De-Finetti's theorem. In addition, the connections among Pólya urn model, exchangeability, the De-Finetti's theorem and the Dirichlet distribution will be discussed. Chapter 4 contains one application of the Dirichlet distribution. It is an unsupervised algorithm given for estimating parameters of a finite mixture model based on the Dirichlet distribution.

Chapter 2

The Dirichlet Distribution

As a multivariate generalization of the Beta distribution, the Dirichlet distribution can also be derived from the Gamma distribution. The definition of the Dirichlet distribution and some basic properties (including statistical and information-theoretic quantities) will be reviewed in this chapter. The method of deriving the moment generating function, entropy, divergence, and mutual information will also be shown.

2.1 The Gamma Distribution

Definition 2.1.1. *A random variable X is said to have a Gamma distribution with parameters α and β if it has a probability density function (pdf) $f(x)$ as shown below*

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}}, & \text{if } 0 < x < \infty \\ 0, & \text{otherwise} \end{cases},$$

where $\alpha > 0$, $\beta > 0$ *and* $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ *is the Gamma function.*

Here, α is a shape parameter and β is a scale parameter for the Gamma density. We denote this distribution by $G(\alpha, \beta)$.

Theorem 2.1.1. *[25] Let* $X_1, \ldots, X_n$ *be independent random variables. Suppose* X_i *has a* $G(\alpha_i, \beta)$ *distribution for* $i = 1, \ldots, n$. *Then* $Y = \sum_{i=1}^n X_i$ *has a* $G(\sum_{i=1}^n \alpha_i, \beta)$ *distribution.*

Suppose we want to generate a random variable from the Gamma distribution with a positve integer valued shape parameter α and scale parameter $\beta > 0$, i.e., $X \sim G(\alpha, \beta)$. Note that the Gamma distribution has the same density function as the exponential distribution when $\alpha = 1$. Therefore, we can generate X by summing the independent and identically distributed (i.i.d.) exponential random variables E_i, where $E_i \sim Exp(\beta)$ and $\beta > 0$ is a real number. $X = \sum_{i=1}^{\alpha} E_i, i = 1, \ldots, \alpha$ and α is an arbitrarily integer. If a random variable U_i is uniformly distributed on $[0, 1]$, then $-\frac{1}{\beta}\log U_i$ is an exponential distribution $Exp(\beta)$ by using the transformation technique [17]. This is one method of generating exponential variables from the uniform distribution. Hence, X can be expressed as

$$X = \sum_{i=1}^{\alpha} E_i = -\frac{1}{\beta}\sum_{i=1}^{\alpha} \log U_i.$$

The above Gamma generator, which is obtained from exponential random variates, is not a good Gamma generator because this strategy increases time linearly with the parameter α [17]. We know that when $0 < \alpha \leq 1$, the Gamma density approaches to infinity at 0. When $\alpha > 1$, the Gamma density

is close to the normal distribution for large values of α. Hence, we consider the problem of generating a Gamma random variable X for $0 < \alpha \leq 1$ and $\alpha > 1$. For arbitrary values of α, there are some good approaches for creating efficient Gamma generators by using rejection algorithms. Classification of the rejection algorithms is dependent on the family of dominating curves used. For example, Ahrens and Dieter (1974 and 1982) for $\alpha > 1$ and Cheng and Feast (1979) for $0 < \alpha \leq 1$ have shown methods of generating a Gamma random variable by using rejection algorithms.

2.2 The Beta Distribution

Definition 2.2.1. *A random variable Y is said to have a Beta distribution with parameters α and β if it has a pdf $f(y)$ as shown below*

$$f(y) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, & \text{if } 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}, \tag{2.1}$$

where $\alpha > 0$, $\beta > 0$.

Note that Beta function is defined by $B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy$.

Let X_1 and X_2 be two independent random variables with Gamma distribution $G(\alpha, 1)$ and $G(\beta, 1)$, respectively. The joint pdf of X_1 and X_2 is

$$f(x_1, x_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} {x_1}^{\alpha-1} {x_2}^{\beta-1} e^{-x_1-x_2},$$

where $0 < x_1 < \infty$, $0 < x_2 < \infty$, $\alpha > 0$, $\beta > 0$.

Let $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{X_1+X_2}$. We will show that $Y_1 \sim G(\alpha+\beta, 1)$ and $Y_2 \sim$ Beta (α, β) [25].

Let the space S be the first quadrant of the x_1x_2-plane. The space S contains points on the coordinate axes. $y_1 = x_1 + x_2$ and $y_2 = x_1/(x_1 + x_2)$ can be written as $x_1 = y_1y_2$, $x_2 = y_1(1 - y_2)$. Thus,

$$J = \begin{vmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{vmatrix} = -y_1$$

and the joint pdf of Y_1 and Y_2 is

$$f(y_1, y_2) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} {y_1}^{\alpha+\beta-1} e^{-y_1} {y_2}^{\alpha-1} (1 - y_2)^{\beta-1},$$

where $0 < y_1 < \infty$, $0 < y_2 < \infty$.

Since Y_1 and Y_2 are independent, the marginal pdf of Y_2 is

$$\begin{aligned} f_2(y_2) &= \frac{{y_2}^{\alpha-1}(1 - y_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty {y_1}^{\alpha+\beta-1} e^{-y_1}\, dy_1 \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} {y_2}^{\alpha-1}(1 - y_2)^{\beta-1}, \quad 0 < y_2 < 1. \end{aligned}$$

Hence, Y_2 has a Beta distribution with parameters α and β.

Also, $f(y_1, y_2) = f_1(y_1)f_2(y_2)$. Then, Y_1 must have pdf

$$f_1(y_1) = \frac{1}{\Gamma(\alpha+\beta)} y_1^{\alpha+\beta-1} e^{-y_1}, \quad 0 < y_1 < \infty,$$

which is $Y_1 \sim G(\alpha+\beta, 1)$.

2.3 Deriving the Dirichlet Distribution

Let X_i be a random variable from the Gamma distribution $G(\alpha_i, 1),\ i = 1, \ldots, k$, and let $X_1, \ldots, X_k$ be independent. The joint pdf of $X_1, \ldots, X_k$ is

$$f(x_1, \ldots, x_k) = \begin{cases} \prod_{i=1}^{k} \frac{1}{\Gamma(\alpha_i)} x_i^{\alpha_i - 1} e^{-x_i}, & \text{if } 0 < x_i < \infty \\ 0, & \text{otherwise} \end{cases}.$$

Let

$$Y_i = \frac{X_i}{X_1 + X_2 + \cdots + X_k}, \quad i = 1, 2, \ldots, k-1$$

and

$$Z_k = X_1 + X_2 + \cdots + X_k.$$

By using the change of variables technique, this transformation maps $M = \{(x_1, \ldots, x_k) : 0 < x_i < \infty,\ i = 1, \ldots, k\}$ onto $N = \{(y_1, \ldots, y_{k-1}, z_k) : y_i > 0,\ i = 1, \ldots, k-1,\ 0 < z_k < \infty,\ y_1 + \cdots + y_{k-1} < 1\}$. The inverse functions are $x_1 = y_1 z_k,\ x_2 = y_2 z_k,\ \ldots,\ x_{k-1} = y_{k-1} z_k,\ x_k = z_k(1 - y_1 - \cdots - y_{k-1})$. Hence, the Jacobian is

$$J = \begin{vmatrix} z_k & 0 & \cdots & 0 & y_1 \\ 0 & z_k & \cdots & 0 & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & z_k & y_{k-1} \\ -z_k & -z_k & \cdots & -z_k & (1 - y_1 - \cdots - y_{k-1}) \end{vmatrix} = z_k^{k-1}.$$

Then, the joint pdf of $Y_1, \ldots, Y_{k-1}, Z_k$ is

$$f(y_1, \ldots, y_{k-1}, z_k) = \frac{y_1^{\alpha_1 - 1} \cdots y_{k-1}^{\alpha_{k-1} - 1} (1 - y_1 - \cdots - y_{k-1})^{\alpha_k - 1}}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} e^{-z_k} z_k^{\alpha_1 + \cdots + \alpha_k - 1}.$$

By integrating out z_k,the joint pdf of $Y_1, \ldots, Y_{k-1}$ is

$$f(y_1, \ldots, y_{k-1}) = \frac{\alpha_1 + \cdots + \alpha_k}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} y_1^{\alpha_1 - 1} \cdots y_{k-1}^{\alpha_{k-1} - 1} (1 - y_1 - \cdots - y_{k-1})^{\alpha_k - 1},$$

where $y_i > 0$, $y_1 + \cdots + y_{k-1} < 1$, $i = 1, \ldots, k-1$. The joint pdf of the random variables $Y_1, \ldots, Y_{k-1}$ is known as the pdf of the Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$. Furthermore, it is clear that Z_k has a Gamma distribution $G(\sum_{i=1}^{k} \alpha_i, 1)$ and Z_k is independent of $Y_1, \ldots, Y_{k-1}$ [25].

2.4 Definition and Properties

Definition 2.4.1. *Let $Y^k = [Y_1, \ldots, Y_k]$ be a vector with k components, where $Y_i \geq 0$ for $i = 1, 2, \ldots, k$ and $\sum_{i=1}^{k} Y_i = 1$. Also, let $\alpha^k = [\alpha_1, \alpha_2, \ldots, \alpha_k]$, where $\alpha_i > 0$ for each i. Then the Dirichlet probability density function is*

$$f(y^k) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} y_i^{\alpha_i - 1},$$

where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$, $y_i > 0$, $y_1 + \cdots + y_{k-1} < 1$ and $y_k = 1 - y_1 - \cdots - y_{k-1}$. We denote this distribution by $Dir(\alpha_1, \alpha_2, \ldots, \alpha_k)$ [34].

The Dirichlet distribution is a distribution with k positive parameters α^k with respect to a k-dimensional space. We observe that if $k = 2$, $f(y_1, y_2)$ is a pdf of the Beta distribution with parameters α_1 and α_2, which is a special case. The probability density function of the Dirichlet distribution for k random variables is a $k-1$ dimensional probability simplex that exists on a k dimensional space.

When each parameter α_i has the same value, it is called the symmetric

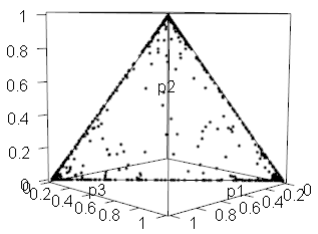
Dirichlet distribution. In this case the density with k components is symmetrically distributed over the $k-1$-dimensional simplex in a k dimensional space.

In Figure 2.1, we plot 1000 points generated from the Dirichlet distribution in a 3-dimensional space with different parameter α^3 values. When $0 < \alpha_1, \alpha_2, \alpha_3 < 1$, the density congregates at the edges of the simplex. Note that in (a) $\alpha^3 = (0.1, 0.1, 0.1)$, the density congregates to at the edges of the triangle. This 2-dimensional simplex represents the sample space of Y_1, Y_2, and Y_3 in 3-dimensional space.

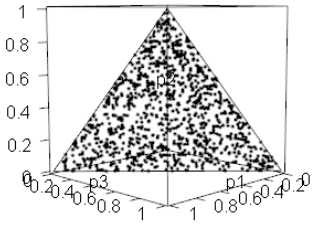
As the value of α^3 increases to $(1, 1, 1)$, the density becomes uniformly distributed over the triangle. When $\alpha_1, \alpha_2, \alpha_3 > 1$, the density becomes more concentrated on the center of the simplex. This is shown in (c) $\alpha^3 = (20, 20, 20)$. In (d), we note that the density plot is not symmetric as the value of $\alpha_1, \alpha_2, \alpha_3$ are not identical.

We next introduce some notations which we will use in the following derivations. We set a random vector $Y^k = [Y_1, \ldots, Y_k]$ to have a Dirichlet distribution with positive parameters $\alpha_1, \ldots, \alpha_k$. It is denoted by $Y^k \sim Dir(\alpha_1, \ldots, \alpha_k)$. Then, for $i = 1, 2, \ldots, k$, $Y_i \geq 0$ and $Y_k = 1 - \sum_{i=1}^{k-1} Y_i$. Also, let $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

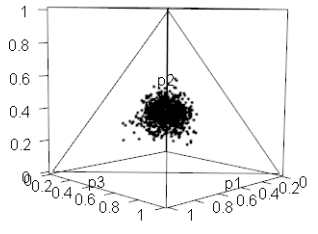
Let X_i be an independent random variable from the Gamma distribution $G(\alpha_i, 1)$ for $i = 1, 2, \ldots, k$, and let $X_1, \ldots, X_k$ be independent. Also, let $Z_k = X_1 + X_2 + \cdots + X_k$; then Z_k has a Gamma distribution $G(\sum_{i=1}^{k} \alpha_i, 1)$.



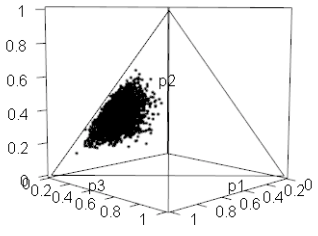
(a) Dir(0.1,0.1,0.1)



(b) Dir(1,1,1)



(c) Dir(20,20,20)



(d) Dir(5,15,25)

Figure 2.1: 1000 points generated from the Dirichlet distribution with parameter $\alpha^3 = (\alpha_1, \alpha_2, \alpha_3)$.

1. **Mean**: $E[Y_i] = \frac{\alpha_i}{\alpha_0}$, $i = 1, 2, \ldots, k$ [34].

   Proof.

$$
\begin{aligned}
E[Y_1] &= \int \cdots \int y_1 \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} y_i^{\alpha_i - 1} dy_1 \cdots dy_k \\
&= \int \cdots \int \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} y_1 y_1^{\alpha_1 - 1} \prod_{i=2}^{k-1} y_i^{\alpha_i - 1} (1 - \sum_{i=1}^{k-1} y_i)^{\alpha_k - 1} dy_1 \cdots dy_{k-1} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \prod_{i=2}^{k} \Gamma(\alpha_i)} \frac{\Gamma(\alpha_1 + 1) \prod_{i=2}^{k} \Gamma(\alpha_i)}{\Gamma(\alpha_0 + 1)} \\
&= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + 1)} \frac{\Gamma(\alpha_1 + 1)}{\Gamma(\alpha_1)} \\
&= \frac{\alpha_1}{\alpha_0}
\end{aligned}
$$

Hence, $E[Y_i] = \frac{\alpha_i}{\alpha_0}$, $i = 1, 2, \ldots, k$. □

2. **Variance**: $VAR(Y_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)}$, $i = 1, 2, \ldots, k$ [34].

Proof. We have shown that $E[Y_i] = \frac{\alpha_i}{\alpha_0}$, $i = 1, 2, \ldots, k$.

Similarly,

$$\begin{aligned} E[Y_i^2] &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + 2)} \frac{\Gamma(\alpha_i + 2)}{\Gamma(\alpha_i)} \\ &= \frac{(\alpha_i + 1)\alpha_i}{(\alpha_0 + 1)\alpha_0}. \end{aligned}$$

Hence,

$$\begin{aligned} VAR(Y_i) &= E[Y_i^2] - E[Y_i]^2 \\ &= \frac{(\alpha_i + 1)\alpha_i}{(\alpha_0 + 1)\alpha_0} - \left(\frac{\alpha_i}{\alpha_0}\right)^2 \\ &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}. \end{aligned}$$

□

3. **Covariance matrix**: $COV(Y_i, Y_j) = \frac{-\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, k$ and $i \neq j$ [34].

Proof. We have shown that $E[Y_i] = \frac{\alpha_i}{\alpha_0}$.

Similarly,

$$\begin{aligned} E[Y_i Y_j] &= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0 + 2)} \frac{\Gamma(\alpha_i + 1)}{\Gamma(\alpha_i)} \frac{\Gamma(\alpha_j + 1)}{\Gamma(\alpha_j)} \\ &= \frac{\alpha_i\alpha_j}{\alpha_0(\alpha_0 + 1)}, \quad i \neq j. \end{aligned}$$

Hence,

$$\begin{aligned} COV(Y_i, Y_j) &= E[Y_i Y_j] - E[Y_i]E[Y_j] \\ &= \frac{\alpha_i \alpha_j}{\alpha_0(\alpha_0+1)} - \frac{\alpha_i}{\alpha_0}\frac{\alpha_j}{\alpha_0} \\ &= \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}, \quad i \neq j. \end{aligned}$$

□

4. **The marginal distribution of** Y_i : Beta $(\alpha_i, \sum_{j=1}^{k} \alpha_i - \alpha_i)$, $i = 1, 2, \ldots, k$ [6].

   Proof. We want to show that the marginal distribution of Y_i is $Beta(\alpha_i, \sum_{j=1}^{k} \alpha_i - \alpha_i)$.

   First, we want to know the distribution of $Z_k - X_i$. By Theorem 2.1.1,

$$Z_k - X_i \sim G\left(\sum_{j=1}^{k} \alpha_i - \alpha_i, 1\right).$$

   Then, Y_i is a Beta distribution

$$\begin{aligned} Y_i &= \frac{X_i}{Z_k} \\ &= \frac{X_i}{X_i + (Z_k - X_i)} \\ &\sim Beta(\alpha_i, \sum_{j=1}^{k} \alpha_i - \alpha_i). \end{aligned}$$

   Therefore, the marginal distribution of Y_i is $Beta(\alpha_i, \sum_{j=1}^{k} \alpha_i - \alpha_i)$ for $i = 1, \ldots, k$ and $0 < Y_i < 1$. □

Note that this marginal distribution is equal to the Dirichlet distribution when $k = 2$. Thus, the Dirichlet distribution is a multivariate generalization of the Beta distribution.

5. **The two-dimensional joint distribution of** (Y_i, Y_j): $Dir(\alpha_i, \alpha_j, \sum_{l=1}^{k} \alpha_i - \alpha_i - \alpha_j)$, $1 \leq i < j \leq k$ [34].

   Proof. Note that $X_i \sim G(\alpha_i, 1)$ and $X_j \sim G(\alpha_j, 1)$.

   First, we want to know the distribution of $Z_k - X_i - X_j$. By Theorem 2.1.1,

   $$Z_k - X_i - X_j \sim G\left(\sum_{l=1}^{k} \alpha_i - \alpha_i - \alpha_j, 1\right).$$

   Then, (Y_i, Y_j) is a Dirichlet distribution

   $$\begin{aligned}(Y_i, Y_j) &= \left(\frac{X_i}{Z_k}, \frac{X_j}{Z_k}\right) \\ &= \left(\frac{X_i}{X_i + X_j + (Z_k - X_i - X_j)}, \frac{X_j}{X_i + X_j + (Z_k - X_i - X_j)}\right) \\ &\sim Dir(\alpha_i, \alpha_j, \sum_{l=1}^{k} \alpha_i - \alpha_i - \alpha_j).\end{aligned}$$

   Therefore, the two-dimensional joint distribution of Y_i, Y_j is $Dir(\alpha_i, \alpha_j, \sum_{l=1}^{k} \alpha_i - \alpha_i - \alpha_j)$ for $1 \leq i < j \leq k$ and $0 < Y_i, Y_j < 1$. □

6. **The conditional joint distribution of** $Y_i^{'} = \frac{Y_i}{1-\sum_{j=1}^{s} Y_j}$ **given** $Y_1 = y_1, \ldots, Y_s = y_s$ ($[Y_i^{'}]_{i=s+1}^{k}$ is independent of $Y_1, \ldots, Y_s$): $Dir(\alpha_{s+1}, \ldots, \alpha_{k-1}, \alpha_k)$, $i = s+1, \ldots, k$ and $0 < s < k$ [34].

   Proof. Let $Y_i^{'} = \frac{Y_i}{1-\sum_{j=1}^{s} Y_j}, \quad i = s+1, \ldots, k.$

We have $Y_i = \frac{X_i}{X_1+X_2+\cdots+X_k}, \quad i = 1, 2, \ldots, k.$

Then,

$$1 - \sum_{j=1}^{s} Y_j = \frac{X_{s+1} + \cdots + X_k}{X_1 + X_2 + \cdots + X_k}.$$

Hence,

$$[Y_i^{'}]_{i=s+1}^{k} = \left[\frac{X_i}{X_{s+1} + \cdots + X_k}\right]_{i=s+1}^{k}$$

Let $Z = X_{s+1}+\cdots+X_k-X_i$, $i = s+1, \ldots, k$. Z is Gamma distributed according to the additive property of the Gamma distribution. Then $Y_i^{'} = \frac{X_i}{X_i+Z}$ is independent of $X_i + Z$, $i = s+1, \ldots, k$ (see deriving the Beta distribution in Section 2.2).

Also,

$$Y_j = \frac{X_j}{X_1 + \cdots + X_k} = \frac{X_j}{X_1 + \cdots + X_s + Z + X_i}, \; j = 1, \ldots, s.$$

Since $Y_i^{'}$ is independent of $\frac{X_j}{X_1+\cdots+X_s}$ and $X_i + Z$, $i = s+1, \ldots, k$, $j = 1, \ldots, s$. Thus, $[Y_i^{'}]_{i=s+1}^{k}$ is independent of Y_j, $j = 1, \ldots, s$.

Therefore, The conditional joint distribution of $Y_i^{'} = \frac{Y_i}{1-\sum_{j=1}^{s} Y_j}$ given $Y_1 = y_1, \ldots, Y_s = y_s$ is $Dir(\alpha_{s+1}, \ldots, \alpha_{k-1}, \alpha_k)$, $i = s+1, \ldots, k$ and $0 < s < k$. □

7. **Product moments** [6]. The product moment with non-negative integers $n_1, \ldots, n_k$ is

$$E(\prod_{i=1}^{k} Y_i^{n_i}) = \int \cdots \int \prod_{i=1}^{k} y_i^{n_i} f(y^k) dy_1 \cdots dy_k$$

$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \int \cdots \int y_1^{n_1+\alpha_1-1}$$
$$\cdots (1 - \sum_{i=1}^{k-1} y_i)^{n_k+\alpha_k-1} dy_1 \cdots dy_{k-1}$$
$$= \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1 + n_1 + \cdots + \alpha_k + n_k)} \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)}.$$

We will use the result of product moments to derive the moment generating function of $Y^k = [Y_1, \ldots, Y_k]$.

8. **Moment generating function**

By using Property (7), we can obtain the moment generating function of $Y^k = [Y_1, \ldots, Y_k]$. Let $t = (t_1, \ldots, t_k)^T \in \mathscr{R}^k$. The moment generating function of Y^k at t is

$$\begin{aligned}
E(e^{t^T Y^k}) &= \int \cdots \int e^{t^T y} f(y^k) dy_1 \cdots dy_k \\
&= \int \cdots \int \sum_{m=0}^{\infty} \frac{(t^T y^k)^m}{m!} f(y^k) dy_1 \cdots dy_k \\
&= \sum_{m=0}^{\infty} \frac{1}{m!} \int \cdots \int (t^T y^k)^m f(y^k) dy_1 \cdots dy_k \\
&\overset{(a)}{=} \sum_{m=0}^{\infty} \frac{1}{m!} \Bigg[\int \cdots \int \sum_{n_1+n_2+\cdots+n_k=m} \frac{m!}{n_1! n_2! \cdots n_k!} \\
&\times \prod_{i=1}^{k} (t_i y_i)^{n_i} f(y^k) dy_1 \cdots dy_k \Bigg] \\
&= \sum_{m=0}^{\infty} \frac{1}{m!} \Bigg[\sum_{n_1+n_2+\cdots+n_k=m} \frac{m!}{n_1! n_2! \cdots n_k!} \\
&\times \Bigg[\prod_{i=1}^{k} (t_i)^{n_i} \int \cdots \int \prod_{i=1}^{k} (y_i)^{n_i} f(y^k) dy_1 \cdots dy_k \Bigg] \Bigg]
\end{aligned}$$

$$= \sum_{m=0}^{\infty} \frac{1}{m!} \Bigg[\sum_{n_1+n_2+\cdots+n_k=m} \frac{m!}{n_1!n_2!\cdots n_k!} \prod_{i=1}^{k} (t_i)^{n_i} E(\prod_{i=1}^{k} Y_i^{n_i}) \Bigg]$$

$$= \sum_{m=0}^{\infty} \frac{1}{m!} \Bigg[\sum_{n_1+n_2+\cdots+n_k=m} \frac{m!}{n_1!n_2!\cdots n_k!} \prod_{i=1}^{k} (t_i)^{n_i}$$

$$\times \left[\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1 + n_1 + \cdots + \alpha_k + n_k)} \prod_{i=1}^{k} \frac{\Gamma(\alpha_i + n_i)}{\Gamma(\alpha_i)} \right] \Bigg].$$

In step (a), we apply the multinomial theorem

$$(x_1 + x_2 + \cdots + x_k)^m = \sum_{n_1+n_2+\cdots+n_k=m} \frac{m!}{n_1!n_2!\cdots n_k!} \prod_{i=1}^{k} x_i^{n_i}$$

for any positive integer k and any non-negative integer m.

9. **Differential entropy** :

$$h(Y^k) = logB(\alpha^k) + (\alpha_0 - k)\psi(\alpha_0) - \sum_{i=1}^{k} (\alpha_i - 1)\psi(\alpha_i),$$

where $\psi(x)$ is the Digamma function and $B(\alpha^k) = \frac{\prod_{i=i}^{k} \Gamma(\alpha_i)}{\Gamma(\alpha_0)}$.

In information theory, entropy is a key concept which was introduced by Claude E. Shannon in 1948 [15]. Entropy is a measure of the uncertainty of a random variable. Differential entropy is the entropy of a continuous random variable.

Definition 2.4.2. *[15] Differential entropy is defined as*

$$h(Y^k) = E[-\log p(Y^k)] = -\int p(y^k) \log p(y^k) dy^k, \quad 0 < h(Y^k) < 1,$$

where $p(y^k)$ is the pdf of the random vector Y^k. Before we derive the differential entropy of the Dirichlet distribution, we first examine the expected value of $\log Y_i$, $i = 1, 2, \ldots, k$. From [32], we have

$$E[\log Y_i] = \psi(\alpha_i) - \psi(\alpha_0),$$

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

Therefore, the differential entropy is

$$\begin{aligned} h(Y^k) &= -E[\log p(y^k)] \\ &= -E\left[\log \frac{1}{B(\alpha)} \prod_{i=1}^{k} y_i^{\alpha_i - 1}\right] \\ &= \log B(\alpha) + \sum_{i=1}^{k} (\alpha_i - 1) E[\log Y_i] \\ &= \log B(\alpha) + \sum_{i=1}^{k} (\alpha_i - 1)(\psi(\alpha_i) - \psi(\alpha_0)) \\ &= \log B(\alpha) + (\alpha_0 - k)\psi(\alpha_0) - \sum_{i=1}^{k} (\alpha_i - 1)\psi(\alpha_i). \end{aligned} \tag{2.2}$$

When $k = 2$ in (2.2), we obtain the special case of the differential entropy for the Beta distribution. Let Y_2 be a random variable from the Beta distribution with parameters α and β (pdf of the Beta distribution is defined as (2.1) in Section (2.2). Thus, the differential entropy of the Beta distribution is

$$h(Y_2) = \log B(\alpha, \beta) + (\alpha + \beta - 2)\psi(\alpha + \beta) - (\alpha - 1)\psi(\alpha) - (\beta - 1)\psi(\beta).$$

The differential entropy of the Beta distribution is also given in [15].

10. **Divergence between two Dirichlet distributions**:

Divergence measures the distance between two distributions f and g. If there exists a random variable X with the true distribution f, then divergence is a measure of the inefficiency when assuming the distribution of X is g [15].

Definition 2.4.3. *The divergence (relative entropy; Kullback-Leibler distance) $D(f(y^k)||g(y^k))$ between two densities $f(y^k)$ and $g(y^k)$ is defined by*

$$D(f(y^k)||g(y^k)) = \int \dots \int f(y^k) log \frac{f(y^k)}{g(y^k)} dy_1 \dots dy_k.$$

Note that $D(f(y^k)||g(y^k)) \geq 0$ since the Kullback–Leibler divergence is always non-negative.

Suppose there are two Dirichlet distributions $f(y^k) \sim Dir(\alpha_1, \dots, \alpha_k)$ and $g(y^k) \sim Dir(\beta_1, \dots, \beta_k)$. Then, the divergence between these two Dirichlet distributions is

$$\begin{aligned} D(f(y^k)||g(y^k)) &= \int \dots \int f(y^k) log \, \frac{f(y^k)}{g(y^k)} \, dy_1 \dots dy_k \\ &= \int \dots \int f(y^k) log \, f(y^k) \, dy_1 \dots dy_k \\ &- \int \dots \int f(y^k) log \, g(y^k) \, dy_1 \dots dy_k \\ &= \int \dots \int f(y^k) \big\{ log \, \Gamma(\sum_{i=1}^{k} \alpha_i) - \sum_{i=1}^{k} log \, \Gamma(\alpha_i) \end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i=1}^{k} (\alpha_i - 1) log\, y_i \big\}\; dy_1 \ldots dy_k \\
&- \int \ldots \int f(y^k) \big\{ log\, \Gamma(\sum_{i=1}^{k} \beta_i) - \sum_{i=1}^{k} log\, \Gamma(\beta_i) \\
&+ \sum_{i=1}^{k} (\beta_i - 1) log\, y_i \big\}\; dy_1 \ldots dy_k \\
&= log\, \Gamma(\sum_{i=1}^{k} \alpha_i) - \sum_{i=1}^{k} log\, \Gamma(\alpha_i) - log\, \Gamma(\sum_{i=1}^{k} \beta_i) + \sum_{i=1}^{k} log\, \Gamma(\beta_i) \\
&+ \sum_{i=1}^{k} (\alpha_i - \beta_i) \int \ldots \int log\, y_i\, f(y^k)\; dy_1 \ldots dy_k, \qquad (2.3)
\end{aligned}
$$

where

$$
\int \ldots \int log\, y_i\, f(y^k)\; dy_1 \ldots dy_k = E[\log Y_i] = \psi(\alpha_i) - \psi(\alpha_0). \qquad (2.4)
$$

Now substitute (2.4) into (2.3). Thus, the divergence between the two Dirichlet distributions is

$$
\begin{aligned}
D(f(y^k)||g(y^k)) &= log\Gamma(\sum_{i=1}^{k} \alpha_i) - \sum_{i=1}^{k} log\Gamma(\alpha_i) - log\Gamma(\sum_{i=1}^{k} \beta_i) \\
&\quad + \sum_{i=1}^{k} log\Gamma(\beta_i) + \sum_{i=1}^{k} (\alpha_i - \beta_i)\, (\psi(\alpha_i) - \psi(\alpha_0)) \\
&= log \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \frac{\prod_{i=1}^{k} \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{k} \beta_i)} \\
&\quad + \sum_{i=1}^{k} (\alpha_i - \beta_i)\, (\psi(\alpha_i) - \psi(\alpha_0))\,. \qquad (2.5)
\end{aligned}
$$

Setting $k = 2$ in (2.5) yields the special case of the divergence between two Beta distributions. Thus, the divergence between two Beta

distributions is

$$D(f_1(y)||g_1(y)) = log\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{\Gamma(\beta_1)\Gamma(\beta_2)}{\Gamma(\beta_1+\beta_2)} + (\alpha_1-\beta_1)\psi(\alpha_1) + (\alpha_2-\beta_2)\psi(\alpha_2) + (\beta_1-\alpha_1+\beta_2-\alpha_2)\psi(\alpha_1+\alpha_2),$$

where $f_1(y) \sim \text{Beta}(\alpha_1, \alpha_2)$ and $f_2(y) \sim \text{Beta}(\beta_1, \beta_2)$, as obtained in [36].

11. **Mutual information**:

Mutual information is a measure of the amount of information shared between two random variables [15].

Definition 2.4.4. *The mutual information, $I(Y_1; Y_2)$, between two continuous random variables Y_1 and Y_2 is defined as*

$$I(Y_1, Y_2) = \int\int f(y_1, y_2) log\frac{f(y_1, y_2)}{f(y_1)f(y_2)} dy_1 dy_2,$$

where $f(y_1)$ and $f(y_2)$ are the marginal probability density functions for the random variable Y_1 and Y_2, respectively. Also, $f(y_1, y_2)$ is the joint probability density function of (Y_1, Y_2).

If $(Y_1, Y_2) \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3)$, then from Properties (4) and (5), we know that the marginal distribution of Y_1 is a Beta distribution with parameters α_1 and β_1 and the marginal distribution of Y_2 is a Beta distribution with parameters α_2 and β_2. Also, the marginal distribution of (Y_1, Y_2) is $Dir(\alpha_1, \alpha_2, \alpha_3)$.

From Definition 2.4.4, the mutual information $I(Y_1, Y_2)$ is equal to

$$I(Y_1, Y_2) = h(Y_1) + h(Y_2) - h(Y_1, Y_2), \tag{2.6}$$

where $h(Y_1)$ is the entropy of the Beta distribution, Beta(α_1, β_1), $h(Y_2)$ is the entropy of the Beta distribution, Beta(α_2, β_2), and $h(Y_1, Y_2)$ the entropy of the Dirichlet distribution, $Dir(\alpha_1, \alpha_2, \alpha_3)$.

The following entropy expressions of the Beta distribution and the Dirichlet distribution can be obtained by using Property (9).

$$\begin{aligned} h(Y_1) = &\log B(\alpha_1, \beta_1) + (\alpha_1 + \beta_1 - 2)\psi(\alpha_1 + \beta_1) \\ &- (\alpha_1 - 1)\psi(\alpha_1) - (\beta_1 - 1)\psi(\beta_1) \end{aligned} \tag{2.7}$$

$$\begin{aligned} h(Y_2) = &\log B(\alpha_2, \beta_2) + (\alpha_2 + \beta_2 - 2)\psi(\alpha_2 + \beta_2) \\ &- (\alpha_2 - 1)\psi(\alpha_2) - (\beta_2 - 1)\psi(\beta_2) \end{aligned} \tag{2.8}$$

$$\begin{aligned} h(Y_1, Y_2) = &\log B(\alpha_1, \alpha_2, \alpha_3) + (\alpha_1 + \alpha_2 + \alpha_3 - 3)\psi(\alpha_1 + \alpha_2 + \alpha_3) \\ &- \sum_{i=1}^{3} (\alpha_i - 1)\psi(\alpha_i) \end{aligned} \tag{2.9}$$

We can thus obtain the mutual information $I(Y_1, Y_2)$ by substituting $h(Y_1)$, $h(Y_2)$, and $h(Y_1, Y_2)$ in (2.6).

The results of the eleven properties of the Dirichlet distribution above are summarized in Table 2.1.

Table 2.1: Properties of the Dirichlet distribution.

Notation	$Y^k \sim Dir(\alpha_1, \ldots, \alpha_k)$
Parameters	$\alpha_1, \ldots, \alpha_k$, $\alpha_i > 0$ for $i = 1, 2, \ldots, k$
Support	$Y_i \geq 0$, $Y_k = 1 - Y_1 - \cdots - Y_{k-1}$ for $i = 1, 2, \ldots, k$ $\sum_{i=1}^{k} Y_i = 1$
Mean	$E[Y_i] = \frac{\alpha_i}{\alpha_0}$, $i = 1, 2, \ldots, k$
Variance	$VAR(Y_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0+1)}$, $i = 1, 2, \ldots, k$
Covariance matrix	$COV(Y_i, Y_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}$, $i, j = 1, 2, \ldots, k$ and $(i \neq j)$
The marginal distribution of Y_i	Beta $(\alpha_i, \sum_{j=1}^{k} \alpha_i - \alpha_i)$, $i = 1, 2, \ldots, k$
The two-dimensional joint distribution of (Y_i, Y_j)	$Dir(\alpha_i, \alpha_j, \sum_{l=1}^{k} \alpha_i - \alpha_i - \alpha_j)$, $1 \leq i < j \leq k$

The conditional joint distribution of $\left[Y_i'\right]_{i=s+1}^{k}$	$Dir(\alpha_{s+1}, \ldots, \alpha_{k-1}, \alpha_k)$, $0 < s < k$
Product moments	$E(\prod_{i=1}^{k} Y_i^{n_i}) = \frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1+n_1+\cdots+\alpha_k+n_k)} \prod_{i=1}^{k} \frac{\Gamma(\alpha_i+n_i)}{\Gamma(\alpha_i)}$, for any $n_1, \ldots n_k \geq 0$
Moment generating function	$\sum_{m=0}^{\infty} \frac{1}{m!} \sum \frac{m!}{n_1!n_2!\cdots n_k!} \prod_{i=1}^{k} (t_i)^{n_i} \frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\Gamma(\alpha_1+n_1+\cdots+\alpha_k+n_k)} \prod_{i=1}^{k} \frac{\Gamma(\alpha_i+n_i)}{\Gamma(\alpha_i)}$
Differential entropy	$h(Y^k) = logB(\alpha^k) + (\alpha_0 - k)\psi(\alpha_0) - \sum_{i=1}^{k}(\alpha_i - 1)\psi(\alpha_i)$
Divergence	$log\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)} \frac{\prod_{i=1}^{k}\Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{k}\beta_i)} + \sum_{i=1}^{k}(\alpha_i - \beta_i)\left(\psi(\alpha_i) - \psi(\alpha_0)\right)$

2.5 Generating Dirichlet Distributed Random Variables

In Section 2.1, we reviewed some methods to generate random variables with a Gamma distribution. Also, random variables with a Beta distribution can be generated from random variables with a Gamma distribution. Since the Dirichlet distribution is a multi-dimensional Beta distribution, the stick-breaking approach [21] can be used for generating Dirichlet random variables.

The general idea is to consider a stick with length 1. First, break the stick into two pieces using an appropriate Beta distribution, and keep one piece of stick. Then, break the remaining stick into two pieces appropriately. Repeat this process until there are k pieces of the stick. This stick-breaking method generates a random vector $(Y_1, \ldots, Y_i, \ldots, Y_k)$, which is distributed as a Dirichlet distribution $Dir(\alpha_1, \ldots, \alpha_k)$, where Y_i is the length of the i^{th} piece of the original stick [21].

Mathematically, we generate a random vector $(Y_1, \ldots, Y_k)$ as follows:

- Simulate a random variate $X_j \sim Beta(\alpha_j, \sum_{i=j+1}^{k} \alpha_i)$, where $j = 1, \ldots, k-1$. When $j = 1$, we have $X_1 \sim Beta(\alpha_1, \sum_{i=2}^{k} \alpha_i)$. The first piece of the stick has length $1 \cdot X_1$, such that the length of the remaining stick is $1 - X_1$. Also, set $Y_1 = X_1$.

- When $j = 2$, we have $X_2 \sim Beta(\alpha_2, \sum_{i=3}^{k} \alpha_i)$. The second piece of the stick has length $(1-X_1)X_2$, such that the length of the remaining stick is $(1-X_1)-(1-X_1)X_2 = (1-X_1)(1-X_2)$. Also, set $Y_2 = (1-X_1)X_2$.

  $\vdots$

- When $j = k-1$, we have $X_{k-1} \sim Beta(\alpha_{k-1}, \alpha_k)$. The $(k-1)^{th}$ piece of the stick has length $X_{k-1} \prod_{j=1}^{k-2}(1-X_j)$, such that the length of the remaining stick is $\prod_{j=1}^{k-1}(1-X_j)$. Also, set $Y_{k-1} = X_{k-1} \prod_{j=1}^{k-2}(1-X_j)$. Note that the k^{th} piece of the stick has length $\prod_{j=1}^{k-1}(1-X_j)$ and set $Y_k = \prod_{j=1}^{k-1}(1-X_j)$. We can conclude that $(Y_1, \ldots, Y_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$.

The stick-breaking approach can be used for generating a Dirichlet random variable because of its "neutrality" [21].

Definition 2.5.1. *Let $Y^k = (Y_1, Y_2, \ldots, Y_k)$ be a random vector, where $Y_i \geq 0$ and $\sum_{i=1}^{k} Y_i = 1$. Y^k is completely neutral if Y_j is independent of the random vector $\frac{1}{1-Y_j} Y^k_{-j}$ for each $j = 1, 2, \ldots, k$, where Y^k_{-j} is the vector Y^k with the j^{th} component removed.*

Lemma 2.5.1. *Let $Y^k \sim Dir(\alpha_1, \ldots, \alpha_k)$. Then Y^k exhibits the above neutrality property.*

Proof. We will show the result for $j = k$. The proof is similar for $j = 1, \ldots, k-1$. Let $Y^k \sim Dir(\alpha_1, \ldots, \alpha_k)$. Also let $Q_i = \frac{Y_i}{1-Y_k}$ for $i = 1, 2, \ldots, k-2$, $Q_{k-1} = 1 - \sum_{i=1}^{k-2} Q_i$, and $Q_k = Y_k$. The transformation T of coordinates between $(q_1, q_2, \ldots, q_{k-2}, q_k)$ and $(y_1, y_2, \ldots, y_{k-2}, y_k)$ is

$$(y_1, y_2, \ldots, y_{k-2}, y_k) = (q_1(1-q_k), q_2(1-q_k), \ldots, q_{k-2}(1-q_k), q_k).$$

Hence, the Jacobian is

$$J = \begin{vmatrix} 1-q_k & 0 & \cdots & 0 & -q_1 \\ 0 & 1-q_k & \cdots & 0 & -q_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1-q_k & -q_{k-2} \\ 0 & 0 & \cdots & 0 & 1 \end{vmatrix} = (1-q_k)^{k-2}.$$

The pdf of $Y_1, Y_2, \ldots, Y_{k-2}, Y_k$ is

$$f(y_1, y_2, \ldots, y_{k-2}, y_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \Big(\prod_{i \neq k-1} y_i^{\alpha_i - 1} \Big) \Big(1 - \sum_{i \neq k-1} y_i\Big)^{\alpha_{k-1} - 1}$$

Then, the joint pdf of $Q_1, Q_2, \ldots, Q_k$ is

$$\begin{aligned} f(q_1, q_2, \ldots, q_k) &= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} (\prod_{i=1}^{k-2} (q_i(1-q_k))^{\alpha_i - 1}) q_k^{\alpha_k - 1} \\ & (1 - \sum_{i=1}^{k-2} q_i(1-q_k) - q_k)^{\alpha_{k-1} - 1} \times (1-q_k)^{k-2} \\ &= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} (\prod_{i=1}^{k-2} (q_i(1-q_k))^{\alpha_i - 1}) q_k^{\alpha_k - 1} \\ & ((1-q_k) q_{k-1})^{\alpha_{k-1} - 1} \times (1-q_k)^{k-2} \\ &= \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} (\prod_{i=1}^{k-1} q_i^{\alpha_i - 1}) q_k^{\alpha_k - 1} (1-q_k)^{\sum_{i=1}^{k-1} \alpha_i - 1}. \end{aligned} \tag{2.10}$$

From (2.10), we have

$$\begin{aligned} f(q_1, q_2, \ldots, q_k) &= \left[\frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Gamma(\alpha_k) \Gamma(\sum_{i=1}^{k-1} \alpha_i)} q_k^{\alpha_k - 1} (1-q_k)^{\sum_{i=1}^{k-1} \alpha_i - 1} \right] \\ & \left[\frac{\Gamma(\sum_{i=1}^{k-1} \alpha_i)}{\prod_{i=1}^{k-1} \Gamma(\alpha_i)} \prod_{i=1}^{k-1} q_i^{\alpha_i - 1} \right] \\ &= f_1(q_k) f_2(q_1, q_2, \ldots, q_{k-1}), \end{aligned} \tag{2.11}$$

where

$$f_1(q_k) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Gamma(\alpha_k) \Gamma(\sum_{i=1}^{k-1} \alpha_i)} q_k^{\alpha_k - 1} (1-q_k)^{\sum_{i=1}^{k-1} \alpha_i - 1} \tag{2.12}$$

is the pdf of a Beta distribution with parameters α_k and $\sum_{i=1}^{k-1} \alpha_i$, and

$$f_2(q_1, q_2, \ldots, q_{k-1}) = \frac{\Gamma(\sum_{i=1}^{k-1} \alpha_i)}{\prod_{i=1}^{k-1} \Gamma(\alpha_i)} \prod_{i=1}^{k-1} q_i^{\alpha_i - 1}. \tag{2.13}$$

is the pdf of a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_{k-1}$.

Note that $f(q_1, q_2, \ldots, q_k)$ can be written as the product of $f_1(q_k)$ and $f_2(q_1, q_2, \ldots, q_{k-1})$. Therefore, Q_k is independent of $(Q_1, Q_2, \ldots, Q_{k-1})$. That is Y_k is independent of $(\frac{Y_1}{1-Y_k}, \frac{Y_2}{1-Y_k}, \ldots, \frac{Y_{k-1}}{1-Y_k})$. □

We observe that the distribution of Y_k is $Beta(\alpha_k, \sum_{i=1}^{k-1} \alpha_i)$. Also, the distribution of $(\frac{Y_1}{1-Y_k}, \frac{Y_2}{1-Y_k}, \ldots, \frac{Y_{k-1}}{1-Y_k})$ is $Dir(\alpha_1, \alpha_2, \ldots, \alpha_{k-1})$. By replacing k with $j = 1, 2, \ldots, k$, we have

$$Y_j \sim Beta(\alpha_j, \sum_{j \neq i} \alpha_j). \tag{2.14}$$

Also,

$$\begin{aligned} f(q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_k | q_j) &= \frac{f(q_1, q_2, \ldots, q_k)}{f_1(q_j)} \\ &= f_2(q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_k) \end{aligned}$$

Hence we have

$$(Q_1, \ldots, Q_{j-1}, Q_{j+1}, \ldots, Q_k | Q_j) \sim Dir(\alpha^k_{-j}),$$

which implies

$$(\frac{Y^k_{-j}}{1 - Y_j} | Y_j) \sim Dir(\alpha^k_{-j})$$

and

$$(Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_k | Y_j) \sim (1 - Y_j) Dir(\alpha^k_{-j}), \tag{2.15}$$

where α^k_{-j} is the vector α^k with the j^{th} component removed and $j = 1, 2, \ldots, k$.

The stick-breaking approach can be proved by using the neutrality property, (2.14) and (2.15) derived above. In order to generate the random Dirichlet vector $(Y_1, \ldots, Y_k) \sim Dir(\alpha_1, \ldots, \alpha_k)$, it is sufficient to first generate the marginal distribution of Y_1 with parameters $(\alpha_1, \ldots, \alpha_k)$, then generate the conditional distribution of $(Y_2, Y_3, \ldots, Y_k|Y_1)$ with parameters $(\alpha_1, \ldots, \alpha_k)$. We can apply this idea recursively to generate k pieces of the stick as follows:

- When $j = 1$: In the stick-breaking approach, we generate the length of the first piece of stick $Y_1 \sim Beta(\alpha_1, \sum_{i=2}^{k} \alpha_i)$. Then the length of the remaining stick is $1 - Y_1$. According to (2.15), we have $(Y_2, Y_3, \ldots, Y_k|Y_1) \sim (1 - Y_1)Dir(\alpha_2, \ldots, \alpha_k)$.

- When $j = 2$: Using the marginal distribution in (2.14), we have $(Y_2|Y_1) \sim (1 - Y_1)Beta(\alpha_2, \sum_{i=3}^{k} \alpha_i)$. Then the length of the remaining stick is $(1 - Y_1)(1 - Y_2)$. According to (2.15), we have $(Y_3, Y_4, \ldots, Y_k|Y_1, Y_2) \sim (1 - Y_1)(1 - Y_2)Dir(\alpha_3, \ldots, \alpha_k)$.

- When $3 \leq j \leq k - 2$: If $j - 1$ pieces of stick have been broken off, then the length of the remaining stick is $\prod_{i=1}^{j-1}(1 - Y_i)$. This is analogous to first generate the marginal distribution of $(Y_1, Y_2, \ldots, Y_{j-1})$, then generate the conditional distribution of $(Y_j, Y_{j+1}, \ldots, Y_k|Y_1, Y_2, \ldots, Y_{j-1})$. From the previous step in the recursion,

$$(Y_j, Y_{j+1}, \ldots, Y_k|Y_1, Y_2 \ldots, Y_{j-1}) \sim (\prod_{i=1}^{j-1}(1 - Y_i))Dir(\alpha_j, \alpha_{j+1}, \ldots, \alpha_k).$$

According to (2.14), we have

$$(Y_j|Y_1, Y_2 \ldots, Y_{j-1}) \sim (\prod_{i=1}^{j-1}(1-Y_i))Beta(\alpha_j, \sum_{i=j+1}^{k}\alpha_i),$$

and from (2.15), we have

$$\begin{aligned}&(Y_{j+1}, Y_{j+2}, \ldots, Y_k|Y_1, Y_2, \ldots, Y_j)\\ &\sim (\prod_{i=1}^{j-1}(1-Y_i))(1-Y_j)Dir(\alpha_{j+1}, \alpha_{j+2}, \ldots, \alpha_k),\end{aligned}$$

which implies

$$\begin{aligned}&(Y_{j+1}, Y_{j+2}, \ldots, Y_k|Y_1, Y_2, \ldots, Y_j)\\ &\sim (\prod_{i=1}^{j}(1-Y_i))Dir(\alpha_{j+1}, \alpha_{j+2}, \ldots, \alpha_k).\end{aligned}$$

- When $j = k-1, k$: We have $(Y_{k-1}, Y_k|Y_1, \ldots, Y_{k-2}) \sim (\prod_{i=1}^{k-2}(1-Y_i))Dir(\alpha_{k-1}, \alpha_k)$. Therefore, we split the remainder of the stick in to two pieces by generating $(Y_{k-1}|Y_1, \ldots, Y_{k-2}) \sim (\prod_{i=1}^{k-2}(1-Y_i))Beta(\alpha_{k-1}, \alpha_k)$ and let Y_k be the remainder.

Chapter 3

The Dirichlet Distribution and Exchangeability

In this chapter, we consider the concept of exchangeability and De-Finetti's theorem. The assumption of exchangeability has strong mathematical implications. Bruno de Finetti (1931) showed that an exchangeable binary sequence is a mixture of i.i.d. Bernoulli sequences. Hewitt and Savage (1995) generalized De-Finetti's theorem into the representation theorem. In Chapter 2, we introduced how to generate random variables with a Dirichlet distribution by using the stick-breaking approach. The next approach that will be reviewed in this chapter is the Pólya urn method [31] [21] [10].

3.1 Exchangeability

Consider an experiment where we flip a coin 10 times. We have 9 heads before we observe the 10th flip. From a subjective perspective, it is reason-

able to assume that the next flip will show heads again. From an objective perspective, the result of the 10th flip is dependent on the probability of getting heads or tails. The idea of exchangeability is that we do not care about the order of heads or tails, but about the number of heads or tails that we have already observed.

Definition 3.1.1. *[4] [9]*

An infinite sequence $\{X_1, X_2, \dots\}$ *of random variables is exchangeable if* $\forall n = 1, 2, \dots$

$$X_1, X_2, \dots, X_n \stackrel{d}{=} X_{i_1}, X_{i_2}, \dots, X_{i_n},^4$$

where $\{i_1, i_2, \dots, i_n\}$ *is any permutation of* $\{1, 2, \dots, n\}$.

3.2 De-Finetti's Theorem

Bruno de Finetti(1931) proved that an exchangeable binary sequence is a mixture of i.i.d. Bernoulli sequences.

Theorem 3.2.1. *[9] [33] A binary sequence* $\{X_1, X_2, \dots\}$ *is exchangeable if and only if there exists a distribution function* F *on* $[0, 1]$ *such that for all* $n \geq 1$

$$p(x_1, \dots, x_n) = \int_0^1 q^{s_n}(1-q)^{n-s_n} dF(q),$$

where $s_n = \sum_{i=1}^n x_i$ *and* $p(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n)$ *is the joint probability mass function of* $(X_1, \dots, X_n)$.

[4] This notation means that $\{X_1, X_2, \dots, X_n\}$ and $\{X_{i_1}, X_{i_2}, \dots, X_{i_n}\}$ have the same distribution.

Here F is the distribution function of the limiting frequency:

$$\bar{X}_\infty = \lim_{n\to\infty} \frac{\sum_i X_i}{n}.$$

In other words, $P(\bar{X}_\infty \leq q) = F(q)$.

Conditioning on the unknown parameter q, $\{X_1, X_2, \dots\}$ is an i.i.d. sequence of Bernoulli random variables with parameter q. The joint sampling distribution is obtained by conditioning on q, such that

$$P(X_1 = x_1, \dots, X_n = x_n|q) = \prod_{i=1}^{n} p(x_i|q) = q^{s_n}(1-q)^{n-s_n},$$

where the parameter q is assigned a prior distribution $F(q)$. Since the joint sampling distribution is a function of q, we refer to it as the likelihood function.

De-Finetti's theorem originally focuses on an exchangeable sequence of binary random variables. Now it is also valid for arbitrary random variables. Hewitt and Savage (1995) developed the representation theorem by generalizing De-Finetti's theorem.

Theorem 3.2.2. *[39] Let $(\mathscr{S}, \mathscr{A}, \mu)$ be a probability space. Also let $(\mathscr{X}, \mathscr{B})$ be a Borel space. Let $X_n : \mathscr{S} \to \mathscr{X}$ be measurable for each n. The sequence $\{X_1, X_2, \dots\}$ is exchangeable if and only if there is a random probability measure $\boldsymbol{P}$ on $(\mathscr{X}, \mathscr{B})$ such that, conditional on $\boldsymbol{P} = P$, $X_1, X_2, \dots$ are i.i.d. with distribution P. Also, if the sequence is exchangeable, then the distribution of $\boldsymbol{P}$ is unique, and $\boldsymbol{P}_n(B)$ converges to $\boldsymbol{P}(B)$ with probability 1 for each $B \in \mathscr{B}$, where $\boldsymbol{P}_n$ is the empirical distribution of $X_1, \dots, X_n$.*

Note that $\mathbf{P}_n(B)$ can be expressed as

$$\mathbf{P}_n(B) = \frac{1}{n}\sum_{i=1}^{n} I_B(X_i), \quad \text{for every } B \in \mathscr{B},$$

where I_B is the indicator function of the set B.

3.3 Pólya Urn model

An urn model is a system of one or more urns containing various types of objects. Those objects are normally represented as balls of different colors. There are various rules and schemes for the urn model. Urn models have many applications in different fields including information and communication theory [2] [3] [5] and [42], image processing [7], economics [38] and biology [38]. A typical Pólya urn model is an urn model with one urn containing different colored balls with a replacement scheme. The following discussion is based on [31] and [33].

3.3.1 Pólya Urn and Exchangeability

Consider an urn that initially has b black balls and w white balls. Let $X_n = 1$ when the ball on the n^{th} draw is black and let $X_n = 0$ when the ball on the n^{th} draw is white. Now we draw one ball randomly from the urn. Next, we return the ball we drew with another ball that shares the same color. Suppose we get one black ball in three draws. The probability of this

event happening is

$$p(1,0,0) = p(1)p(0|1)p(0|0,1) = \frac{b}{w+b}\frac{w}{w+b+1}\frac{w+1}{w+b+2},$$

$$p(0,1,0) = p(0)p(1|0)p(0|1,0) = \frac{w}{w+b}\frac{b}{w+b+1}\frac{w+1}{w+b+2},$$

$$p(0,0,1) = p(0)P(0|0)p(1|0,0) = \frac{w}{w+b}\frac{w+1}{w+b+1}\frac{b}{w+b+2},$$

$$p(1,0,0) = p(0,1,0) = p(0,0,1),$$

where b is the number of black balls initially contained in the urn and w is the number of white balls initially contained in the urn.

Note that the probability is only dependent on the number of black balls. Therefore, the probability of any finite sequence of events only depends on observing the number of white or black balls. We can say that the constructed binary sequence $\{X_1, X_2, \dots\}$ from this Pólya urn model is exchangeable, but not independent [31].

3.3.2 Pólya Urn and De-Finetti's Theorem

By using De-Finetti's theorem, we will show that the limiting distribution for the Pólya urn model is

$$\bar{X}_\infty \sim Beta\left(\frac{b}{b+w}, \frac{w}{b+w}\right),$$

where b is the number of black balls initially contained in the urn and w is the number of white balls initially contained in the urn. Conditioning on $\bar{X}_\infty = q$, $X_1, X_2, \ldots$ are independent and Bernoulli distributed with parameter q.

Let $X_n = 1$ when the ball on the n^{th} draw is black and let $X_n = 0$ when the ball on the n^{th} draw is white. The probability of this event happening is

$$\begin{aligned} &P(X_1 = 1, X_2 = 1, \ldots, X_k = 1, X_{k+1} = 0, \ldots, X_n = 0) \\ &\quad = \frac{b(b+1)\cdots(b+k-1)w(w+1)\cdots(w+n-k-1)}{(b+w)(b+w+1)\cdots(b+w+n-1)}. \end{aligned}$$

Let $\sum_{i=1}^{n} X_i = k$ when we get a total of k black balls in n draws. The probability of getting a total of k black balls in n draws is only dependent on the number of black balls. Therefore, the probability of observing a total of k black balls in n draws is

$$\begin{aligned} P(\sum_{i=1}^{n} X_i = k) &= \binom{n}{k} \frac{b(b+1)\cdots(b+k-1)w(w+1)\cdots(w+n-k-1)}{(b+w)(b+w+1)\cdots(b+w+n-1)} \\ &\overset{(1)}{=} \binom{n}{k} \frac{\frac{\Gamma(b+k)}{\Gamma(b)}\frac{\Gamma(w+n-k)}{\Gamma(w)}}{\frac{\Gamma(b+w+n)}{\Gamma(b+w)}} \\ &= \binom{n}{k} \frac{\frac{\Gamma(b+k)\Gamma(w+n-k)}{\Gamma(b+w+n)}}{\frac{\Gamma(b)\Gamma(w)}{\Gamma(b+w)}} \\ &= \binom{n}{k} \frac{\beta(b+k, w+n-k)}{\beta(b,w)}, \end{aligned} \tag{3.1}$$

where $\beta(b,w) = \frac{\Gamma(b)\Gamma(w)}{\Gamma(b+w)}$ is the Beta function. In (1), we use the property of the Gamma function

$$\Gamma(x+1) = x\Gamma(x).$$

We know that the probability of observing a total of k black balls in n

draws conditioning on the parameter q is

$$P(\sum_{i=1}^{n} X_i = k|q) = \binom{n}{k} q^k (1-q)^{n-k}.$$

If the parameter q has a distribution function $F(q)$, then

$$\begin{aligned} P(\sum_{i=1}^{n} X_i = k) &= \int_0^1 P(X'_n = k|q) dF(q) \\ &= \int_0^1 \binom{n}{k} q^k (1-q)^{n-k} dF(q). \end{aligned} \tag{3.2}$$

Since (3.1) and (3.2) are equivalent, we obtain

$$\begin{aligned} \int_0^1 q^k (1-q)^{n-k} dF(q) &= \frac{\beta(b+k, w+n-k)}{\beta(b,w)} \\ &\overset{(3)}{=} \int_0^1 \frac{q^{b+k-1}(1-q)^{w+n-k-1}}{\beta(b,w)} dq \\ &= \int_0^1 q^k (1-q)^{n-k} \frac{q^{b-1}(1-q)^{w-1}}{\beta(b,w)} dq. \end{aligned}$$

Hence,

$$dF(q) = \frac{q^{b-1}(1-q)^{w-1}}{\beta(b,w)} dq.$$

In step (3), we have used the definition of the Beta function, $\beta(x,y) = \int_0^1 q^{x-1}(1-q)^{y-1} dq$. Therefore, we can conclude that $\bar{X}_\infty \sim Beta\left(\frac{b}{b+w}, \frac{w}{b+w}\right)$.

3.4 Pólya urn and the Dirichlet distribution

Previously, we discussed a Pólya urn model with two different colored balls. The constructed binary sequence $\{X_1, X_2, \dots\}$ from this Pólya urn

model is exchangeable. Also, De-Finetti's theorem indicates that this exchangeable binary sequence is a mixture of independent and identically distributed (i.i.d.) Bernoulli sequences.

Now suppose there is an urn that contains k different colored balls. Originally, there are α_i balls of color i, $i = 1, 2, \ldots, k$, and $\alpha_i > 0$. Here, the value of α_i could be any integer which is greater than zero. We draw one ball randomly from the urn and we return the ball we drew with another ball that shares the same color. Repeat this process indefinitely and generate a sequence of balls with colors $(X_1, X_2, \ldots)$. The proportion of different colored balls in the urn will converge to the Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$, as n approaches infinity [21].

The probability of each draw can be shown by the following:

- Randomly draw a ball from the urn with color i. Let X_m be the ball in m^{th} draw.
- First draw: $P(X_1 = i) = \frac{\alpha_i}{\sum_{i=1}^{k} \alpha_i}$.
- Second draw: $P(X_2 = i|X_1) = \frac{\alpha_i + \delta_i(X_1)}{1 + \sum_{i=1}^{k} \alpha_i}$.

  $\vdots$

- m^{th} draw: $P(X_m = i|X_1, \ldots, X_{m-1}) = \frac{\alpha_i + \sum_{j=1}^{m-1} \delta_i(X_j)}{m-1+\sum_{i=1}^{k} \alpha_i}$.

Where $\delta_i(X_j)$ is the indicator function and $\delta_i(X_j) = \begin{cases} 1, \text{ if } X_j = i \\ 0, \text{ if } X_j \neq i \end{cases}$ for $i = 1, \ldots, k$, $j = 1, \ldots, m-1$.

Let Y_i be the total number of color i balls when the urn contains n balls. Also let $\frac{Y_i}{n}$ be the proportion of different colored balls for $i = 1, 2, \ldots, k$. The

random vector $[\frac{Y_1}{n}, \ldots, \frac{Y_k}{n}] \xrightarrow{\mathrm{d}} Y^k \sim Dir(\alpha_1, \ldots, \alpha_k)$ as $n \to \infty$, where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution.

If we do not know the total number of "colors" in the data beforehand, then we need to extend the Pólya urn scheme from a fixed k colors to infinitely many colors. In reference [10], the Pólya urn scheme is extended to a continuum of colors. After n draws , the distribution of colors converges to a limiting discrete distribution μ^* as n approaches infinity and μ^* has a Ferguson distribution with parameter μ.

The Ferguson distribution is described as follows [10]. Let $\mathscr{X}$ be a complete separable metric space. Let μ be any finite positive measure on $\mathscr{X}$. Also, let μ^* be a random probability measure on $\mathscr{X}$. For every finite partition $(B_1, \ldots, B_r)$ of $\mathscr{X}$, if the vector $[\mu^*(B_1), \ldots, \mu^*(B_r)]$ has a Dirichlet distribution with parameter $\mu(B_1), \ldots, \mu(B_r)$, then μ^* has a Ferguson distribution with parameter μ. Note that when $\mu(B_i) = 0$, $\mu^*(B_i) = 0$ with probability 1 [10].

Now we introduce the definition of a Pólya sequence. We say a sequence $\{X_n, n \geq 1\}$ of random variables with values in $\mathscr{X}$ is a Pólya sequence with parameter μ if

$$P(X_1 \in B) = \mu(B)/\mu(\mathscr{X}),\ for\ every\ B \subset \mathscr{X}$$

and

$$P(X_{n+1} \in B | X_1, \ldots, X_n) = \mu_n(B)/\mu_n(\mathscr{X}),\ for\ every\ B \subset \mathscr{X},$$

where $\mu_n = \mu + \sum_{i=1}^{n} \delta(X_i)$ and $\delta(x)$ denotes the unit measure concentrating at x. For finite $\mathscr{X}$, the sequence $\{X_n\}$ represents the results of successive draws from an urn where initially the urn has $\mu(x)$ balls of color x. Then after each draw, the ball drawn is returned to the urn and one additional ball of its same color is also placed within the urn. Note that, without the restriction to finite $\mathscr{X}$, for any function ϕ on $\mathscr{X}$, the sequence $\{\phi(X_n)\}$ is a Pólya sequence with parameter $\phi\mu$, where $\phi\mu(A) = \mu\phi \in A$ [10].

The connections between Pólya sequences and Ferguson distributions with the following theorem are described in [10].

Theorem 3.4.1. *Let $\{X_n\}$ be a Pólya sequence with parameter μ. Then*
(a) $m_n = \frac{\mu_n}{\mu_n(\mathscr{X})}$ converges with probability 1 as $n \to \infty$ to a limiting discrete measure μ^*,
(b) μ^* has a Ferguson distribution with parameter μ and
(c) given μ^*, the variables $X_1, X_2, \ldots$ are independent with distribution μ^*.

This theorem states $[m_n(B_1), \ldots, m_n(B_r)]$ converges to $[\mu^*(B_1), \ldots, \mu^*(B_r)]$ and $[\mu^*(B_1), \ldots, \mu^*(B_r)]$ has a Dirichlet distribution with parameter $\mu(B_1), \ldots, \mu(B_r)$. Then μ^* has a Ferguson distribution with parameter μ.

3.5 Conjugate Prior for the Multinomial Distribution

The Bayesian approach provides decision making under uncertainty, while the prior distribution gives additional information about uncertainty.

Statistical inference can be made by using data and prior information. In the Bayesian approach, incoming data is required to update one's belief. However, the limit of a parametric model in the Bayesian approach is the fixed size of parameters. In a non-parametric model, there are infinite-dimensional parameter spaces, which can be adapted into complex models with large amounts of data. The Dirichlet distribution is a popular conjugate prior for the Multinomial distribution.

The Multinomial distribution is a multivariate generalization of the binomial distribution. The probability mass function is given by

$$f(z_1, \ldots, z_k | y^k = (y_1, y_2, \ldots, y_k)) = \frac{n!}{z_1! z_2! \cdots z_k!} \prod_{i=1}^{k} y_i^{z_i},$$

where $z_i \in \{0, 1, \ldots, n\}$, $i = 1, \ldots, k$, $z_k = n - (z_1 + z_2 + \cdots + z_{k-1})$ and y_i is the probability parameter, $i = 1, \ldots, k$.[21].

The Dirichlet distribution is a conjugate prior distribution for the Multinomial distribution. Note that if the posterior distribution has the same family of distribution as the prior distribution, we say that this prior distribution is a conjugate prior.

Lemma 3.5.1. *If a random vector $Y^k = (Y_1, Y_2, \ldots, Y_k)$ has a Dirichlet distribution $Dir(\alpha_1, \ldots, \alpha_k)$ and $(Z^k | Y^k)$ has a Multinomial distribution, where $Z^k = (Z_1, Z_2, \ldots, Z_k)$. Then the joint posterior $(Y^k | Z^k)$ has a Dirichlet distribution $Dir(\alpha_1 + z_1, \ldots, \alpha_k + z_k)$ [21].*

Proof. Let $f(y^k)$ and $f(y^k | z^k)$ be the pdf of the prior distribution and the posterior distribution respectively. By using the Baye's rule, the posterior

distribution is

$$
\begin{aligned}
f(y^k|z^k) &\propto f(z^k|y^k)f(y^k) \\
&= \left(\frac{n!}{z_1!z_2!\cdots z_k!}\prod_{i=1}^{k} y_i^{z_i}\right)\left(\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k} y_i^{\alpha_i-1}\right) \qquad (3.3) \\
&\propto \prod_{i=1}^{k} y_i^{\alpha_i+z_i-1} \sim Dir(\alpha_1+z_1,\ldots,\alpha_k+z_k). \qquad (3.4)
\end{aligned}
$$

Therefore, the posterior distribution $f(y^k|z^k)$ is also a Dirichlet distribution.

Note that $\propto$ is the proportional symbol. Given $\frac{n!}{z_1!z_2!\cdots z_k!}$ and $\frac{\Gamma(\alpha_1+\cdots+\alpha_k)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}$ are both constant, (3.3) is proportional to (3.4). □

In Bayesian analysis, the Dirichlet distribution can be used as a prior distribution due to its conjugacy. We will review the generalized Dirichlet distribution which is also the conjugate prior distribution for the Multinomial distribution.

Definition 3.5.1. *[41] Let $X^k = (X_1,\ldots,X_k)$ be a random vector that has the generalized Dirichlet distribution $GD(\alpha_1,\alpha_2,\ldots,\alpha_k;\beta_1,\beta_2,\ldots,\beta_k)$ with the density function*

$$f(x^k) = \prod_{i=1}^{k}\frac{1}{B(\alpha_i,\beta_i)}x_i^{\alpha_i-1}(1-x_1-\cdots-x_i)^{\gamma_i},$$

where $\alpha_i > 0$, $\beta_i > 0$, $\gamma_i = \beta_i - \alpha_{i+1} - \beta_{i+1}$ for $i = 1,2,\ldots,k-1$ and $\gamma_k = \beta_k - 1$. Also, $X_1 + X_2 + \cdots + X_k \leq 1$ and $X_j \geq 0$ for $j = 1,2,\ldots,k$.

Note that if we set $\gamma_1 = \gamma_2 = \cdots = \gamma_{k-1} = 0$, we can obtain the density of a Dirichlet distribution. In [14], the derivation and formulae for

$E(X_i), VAR(X_i)$ and $COV(X_r, X_s)$ are shown. They are summarized as follows:

- $E[X_i] = \left(\prod_{j=1}^{i-1} \frac{\beta_j}{\alpha_j+\beta_j}\right) \frac{\alpha_i}{\alpha_i+\beta_i}, \quad i = 1, 2, \ldots, k.$
- $VAR(X_i) = E[X_i]\{\frac{\alpha_i+1}{\alpha_i+\beta_i+1} \prod_{j=1}^{i-1} \frac{\beta_j+1}{\alpha_j+\beta_j+1} - E[X_i]\}, \quad i = 1, 2, \ldots, k.$
- $COV(X_r, X_s) = E[X_s]\{\frac{\alpha_r}{\alpha_r+\beta_r+1} \prod_{j=1}^{r-1} \frac{\beta_j+1}{\alpha_j+\beta_j+1} - E[X_r]\}$, $r = 1, 2, \ldots, k-1$, and $s = r+1, \ldots, k$.

Lemma 3.5.2. *Suppose the random vector $X^k = (X_1, X_2, \ldots, X_k)$ has a generalized Dirichlet distribution $GD(\alpha_1, \alpha_2, \ldots, \alpha_k;\ \beta_1, \beta_2, \ldots, \beta_k)$ and $(Z^{k+1}|X^k)$ has a Multinomial distribution, where $Z^{k+1} = (Z_1, Z_2, \ldots, Z_{k+1})$, $Z_j = z_j$ for $j = 1, 2, \ldots, k+1$, and $z_{k+1} = n - z_1 - z_2 - \cdots - z_k$. Then, the joint posterior $(X^k|Z^{k+1})$ has a generalized Dirichlet distribution $GD(\alpha'_1, \alpha'_2, \ldots, \alpha'_k;\ \beta'_1, \beta'_2, \ldots, \beta'_k)$, where $\alpha'_j = \alpha_j + z_j$ and $\beta'_j = \beta_j + z_{j+1} + z_{j+2} + \cdots + z_{k+1}$ for $j = 1, 2, \ldots, k$[41].*

Proof. Let $f(x^k)$ and $f(x^k|z^{k+1})$ be the pdf of the prior distribution and the posterior distribution respectively. By using Bayes' rule, the posterior distribution is

$$\begin{aligned} f(x^k|z^{k+1}) &\propto f(z^{k+1}|x^k) f(x^k) \\ &= x_1^{(\alpha_1+z_1)-1} x_2^{(\alpha_2+z_2)-1} \cdots x_k^{(\alpha_k+z_k)-1} \\ &\times (1-x_1)^{(\beta_1+z_2+\cdots+z_{k+1})-(\alpha_2+z_2)-(\beta_2+z_3+\cdots+z_{k+1})} \\ &\times (1-x_1-x_2)^{(\beta_2+z_3+\cdots+z_{k+1})-(\alpha_3+z_3)-(\beta_3+z_4+\cdots+z_{k+1})} \\ &\cdots (1-x_1-\cdots-x_k)^{(\beta_k+z_{k+1})-1}. \end{aligned}$$

Let $\alpha'_j = \alpha_j + z_j$ and $\beta'_j = \beta_j + z_{j+1} + z_{j+2} + \cdots + z_{k+1}$ for $j = 1, 2, \ldots, k$. So, the posterior density of $(X^k|Z^{k+1})$ will be

$$\begin{aligned} f(x^k|z^{k+1}) &\propto x_1^{\alpha'_1-1} x_2^{\alpha'_2-1} \cdots x_k^{\alpha'_k-1} \\ &\quad \times (1-x_1)^{\beta'_1-\alpha'_2-\beta'_2} (1-x_1-x_2)^{\beta'_2-\alpha'_3-\beta'_3} \\ &\quad \cdots (1-x_1-\cdots-x_k)^{\beta'_k-1} \\ &= \prod_{j=1}^{k} x_i^{\alpha'_j-1} (1-x_1-\cdots-x_i)^{\gamma'_j}, \end{aligned}$$

where $\gamma'_j = \beta'_j - \alpha'_{j+1} - \beta'_{j+1}$ for $j = 1, 2, \ldots, k-1$ and $\gamma'_k = \beta'_k - 1$. Therefore, the posterior distribution $f(x^k|z^{k+1})$ is also a generalized Dirichlet distribution. □

We have shown that the generalized Dirichlet distribution and the Dirichlet distribution can both be the conjugate prior of the Multinomial distribution. In [28], applications of the Dirichlet distribution to forensic match probabilities are discussed. If data from several distinct populations are available, a Dirichlet conjugate prior can be used for the allele frequency estimation in each of the separate populations.

Chapter 4

Application of the Dirichlet Distribution

Machine learning is an area of study that focuses on the research and creation of computer programs that teach themselves and adapt to new data. Furthermore, machine learning is a cross-section of statistics, computer science, engineering and many other fields. Supervised learning and unsupervised learning are two branches of machine learning. In supervised learning, the machine receives an input and is given a corresponding output. The goal is to predict the correct outcome given new information. In unsupervised learning, the machine receives input measures without outcome measures. The goal is to find patterns within the input measures. Cluster analysis, principal component analysis, and independent component analysis are classical examples of unsupervised learning. Techniques in unsupervised learning are detailed in [22] and [20].

Finite mixtures are statistical models with a variety of applications for

multivariate data such as pattern recognition, computer vision, and image analysis. Finite mixture models can be used to model data from several populations with varying proportions. Also, finite mixture models can demonstrate a process for unsupervised learning such as clustering, as well as represent arbitrarily complex probability functions. Two important issues in finite mixture models are determining the appropriate number of components in a mixture and estimating the parameters of the component of a mixture. Maximum likelihood (ML) estimates of the mixture parameters can be obtained by the expectation-maximization (EM) algorithm [19]. The following discussion is a summary of a Dirichlet mixture model and its application based on [12].

4.1 Dirichlet Mixture

There are various research papers focused on the unsupervised learning of finite mixture models with respect to the Dirichlet distribution. A finite generalized Dirichlet distribution mixture model is used to solve the problem of high-dimensional unsupervised learning [11]. Also, the use of a hybrid stochastic expectation maximization algorithm in estimating the parameters of the generalized Dirichlet distribution is introduced in [11]. Unsupervised learning of finite generalized Dirichlet mixture models for data clustering and feature weighting is also described in [26]. In [43], a recursive algorithm is introduced to select the number of components and estimate the parameters of components in finite mixture models. The Dirichelt prior is used in the solution of the Minimum Message Length model selection criterion. An un-

supervised learning method for human action categories is presented in [35]. Given a video sequence with multiple motions, their algorithm can recognize and localize multiple actions. The unsupervised learning of topic-based clusters from text documents is discussed in [8]. Three batch topic models such as LDA, Dirichlet Compound Multinomial and von-Mises Fisher are discussed in document clustering.

In this chapter, we review an unsupervised learning algorithm for a finite mixture model with multivariate data. This mixture model is based specifically on the Dirichlet distribution [12].

Let $X^k = (X_1, X_2, \ldots, X_k)$ be a random vector with a Dirichlet distribution $Dir(\alpha^k)$, where $\alpha^k = [\alpha_1, \alpha_2, \ldots, \alpha_k]$. Recall that the pdf of the Dirichlet distribution is

$$p(x^k) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}.$$

Note that $\alpha_0 = \sum_{i=1}^{k} \alpha_i$, $x_i > 0$, $x_1 + \cdots + x_{k-1} < 1$ and $x_k = 1 - x_1 - \cdots - x_{k-1}$.

The pdf of a Dirichlet mixture with m components is defined as

$$p(x^k|\Theta) = \sum_{j=1}^{m} p(x^k|\Theta_j)p(j), \tag{4.1}$$

where $p(x^k|\Theta_j)$ is the pdf of the Dirichlet distribution of the j^{th} component for $j = 1, 2, \ldots, m$. Denote $\Theta_j = \alpha_j^k$ as the set of parameters of the j^{th} component for $j = 1, 2, \ldots, m$ and Θ is the complete set of parameters of the

mixture model

$$\Theta = \{\alpha_1^k, \alpha_2^k, \ldots, \alpha_m^k, p(1), p(2), \ldots, p(m)\}.$$

Also, $p(1), p(2), \ldots, p(m)$ are the mixing probabilities that satisfy

$$p(j) \geq 0 \quad and \quad \sum_{j=1}^{m} p(j) = 1, \quad j = 1, 2, \ldots, m.$$

4.2 Maximum Likelihood Estimation

The most common method to estimate the parameters of a mixture model is ML estimation. The expectation maximization (EM) algorithm is an iterative method used in finding ML estimates of parameters. In [16], the EM algorithm was first proposed for estimating the ML estimator (MLE) of stochastic models. A drawback of the EM algorithm is the number of components is required to specify each time. We can use some criterion functions to overcome this problem. Akaike information criterion (AIC) [1], Schwartz's Bayesian information criterion (BIC) [1] and minimum description length (MDL) [24] are three criteria used in [12]. The ML estimation method concerns choosing parameters to maximize the joint density function of the sample (likelihood function). Therefore, we consider

$$\max_{\Theta} p(x^k|\Theta)$$

with constraints $\sum_{j=1}^{m} p(j) = 1$ and $p(j) > 0$ for $j = 1, 2, \ldots, m$. We can consider $p(j)$ as prior probabilities under these constraints.

Now suppose we have a sample that contains n random vectors X_i^k, which are i.i.d., $i = 1, \ldots, n$. We maximize the following function with respect to Θ and Λ

$$\begin{aligned}\Phi(x^k, \Theta, \Lambda) = &\sum_{i=1}^{n} ln(\sum_{j=1}^{m} p(x_i^k|\Theta_j)p(j)) \\ &+ \Lambda(1 - \sum_{j=1}^{m} p(j)) \\ &+ \mu \sum_{j=1}^{m} p(j)ln(p(j)). \end{aligned} \tag{4.2}$$

The first term of (4.2) is the log-likelihood function. Λ is the Lagrange multiplier in the second term. In the last term of (4.2) we use an entropy-based criterion [29]. Also, μ is the ratio of the first term to the last term in (4.2) of each iteration t [12]

$$\mu(t) = \frac{\sum_{i=1}^{n} ln(\sum_{j=1}^{m} p^{(t-1)}(x_i^k|\Theta_j)p^{(t-1)}(j))}{\sum_{j=1}^{m} p^{(t-1)}(j)ln(p^{(t-1)}(j))}. \tag{4.3}$$

In order to optimize (4.2), we need to solve the following equations:

$$\frac{\partial}{\partial \Theta}\Phi(x^k, \Theta, \Lambda) = 0 \tag{4.4}$$

$$\frac{\partial}{\partial \Lambda}\Phi(x^k, \Theta, \Lambda) = 0 \tag{4.5}$$

It is shown in [12] that the estimator of the prior probability $p(j)$ is

$$p(j)^{new} = \frac{\sum_{i=1}^{n} p^{old}(j|x_i^k, \Theta_j) + \mu[p(j)^{old}(1 + ln\ p(j)^{old})]}{n + \mu \sum_{j=1}^{m} p(j)^{old}(1 + ln\ p(j)^{old})},\ j = 1, 2, \ldots, m. \tag{4.6}$$

Note that μ is defined by (4.3) and $p(j|x_i^k, \Theta_j)$ is the a posteriori probability where

$$p(j|x_i^k, \Theta_j) = \frac{p(x_i^k, \Theta_j)p(j)}{p(x_i^k, \Theta)},\ i = 1, \ldots, n,\ j = 1, 2, \ldots, m.$$

Now we want to estimate the parameters $\alpha_j^k,\ j = 1, 2, \ldots, m$. In [12], the Fisher scoring method [30] is used to find these estimates. Denote α_{jl} as one element of the parameter vector α_j^k for each component j, $l = 1, \ldots, k$, $j = 1, \ldots, m$. The derivative of $\Phi(x^k, \Theta, \Lambda)$ with respect to α_{jl} is

$$\begin{aligned}\frac{\partial}{\partial \alpha_{jl}}\Phi(x^k, \Theta, \Lambda) = &\sum_{i=1}^{n} p(j|x_i^k, \alpha_j^k)(ln\, x_{il}) \\ &+ [\psi(\alpha_{0_j}) - \psi(\alpha_{jl})]\sum_{i=1}^{n} p(j|x_i^k, \alpha_j^k), \\ &l = 1, \ldots, k, \\ &j = 1, \ldots, m, \end{aligned} \tag{4.7}$$

where $\psi(\cdot)$ is the Digamma function. However, α_{jl} can become negative during iterations. In order to keep α_{jl} strictly positive, the author of [12] sets $\alpha_{jl} = e^{\beta_{jl}}$. β_{jl} is any real number. Then, the derivative of $\Phi(x^k, \Theta, \Lambda)$ with respect to β_{jl} is

$$\begin{aligned}\frac{\partial}{\partial \beta_{jl}}\Phi(x^k, \Theta, \Lambda) = \alpha_{jl}[&\sum_{i=1}^{n} p(j|x_i^k, \alpha_j^k)(ln\, x_{il}) \\ &+ [\psi(\alpha_{0_j}) - \psi(\alpha_{jl})]\sum_{i=1}^{n} p(j|x_i^k, \alpha_j^k)], \\ &l = 1, \ldots, k, \\ &j = 1, \ldots, m. \end{aligned} \tag{4.8}$$

By using the iterative scheme of the Fisher scoring method, we obtain

$$\begin{aligned}\begin{pmatrix}\hat{\beta}_{jl}\\ \vdots \\ \hat{\beta}_{jk}\end{pmatrix}^{new} &= \begin{pmatrix}\hat{\beta}_{jl}\\ \vdots \\ \hat{\beta}_{jk}\end{pmatrix}^{old} \\ &+ \begin{pmatrix} VAR(\hat{\beta}_{jl}) & \cdots & COV(\hat{\beta}_{jl},\hat{\beta}_{jk}) \\ \vdots & \ddots & \vdots \\ COV(\hat{\beta}_{jk},\hat{\beta}_{j1}) & \cdots & VAR(\hat{\beta}_{jk}) \end{pmatrix}^{old} \\ &\times \begin{pmatrix} \frac{\partial}{\partial\hat{\beta}_{jl}}\Phi(x^k,\Theta,\Lambda) \\ \vdots \\ \frac{\partial}{\partial\hat{\beta}_{jk}}\Phi(x^k,\Theta,\Lambda) \end{pmatrix}^{old}, \quad j=1,\ldots,m. \end{aligned} \tag{4.9}$$

Note that the variance-covariance matrix is obtained by the inverse of the Fisher information matrix $\mathbf{I}$ and

$$\mathbf{I} = -E\left[\frac{\partial^2}{\partial\beta_{jl_1}\partial\beta_{jl_2}}\Phi(x^k,\Theta,\Lambda)\right].$$

4.3 Initialization Algorithm and Dirichlet Mixture Estimation Algorithm

Our goal is to estimate the parameters of a Dirichlet mixture model. Reference [12] presents an algorithm of initializing the parameters and a complete estimation algorithm. The fuzzy C means method [18] and method of moments are used in the initialization algorithm. For each component j

$(j = 1, \ldots, m)$, each element of the vector of parameters α_j^k is defined by

$$\alpha_{jl} = \frac{(x'_{11} - x'_{21})x'_{1l}}{x'_{21} - (x'_{11})^2}, \quad l = 1, \ldots, k-1 \tag{4.10}$$

and

$$\alpha_{jk} = \frac{(x'_{11} - x'_{21})(1 - \sum_{l=1}^{k-1} x'_{1l})}{x'_{21} - (x'_{11})^2}. \tag{4.11}$$

Note that

$$x'_{1l} = \frac{1}{n}\sum_{i=1}^{n} x_{il},$$

$$x'_{2l} = \frac{1}{n}\sum_{i=1}^{n} x_{il}^2, \quad l = 1, \ldots, k, \quad i = 1, \ldots, k$$

and x_{il} is one element of the vector x_i^k which is obtained from the sample data.

Therefore, the **Initialization Algorithm** [12] is as follows:

1. Obtain the elements, mean and covariance matrix of each component j by using the fuzzy C means method, $j = 1, \ldots, m$.

2. Obtain the vector of parameters α_j^k of each component j by using method of moments.

3. Assign the data to clusters.

4. Update $p(j)$ with $p(j) = \frac{\text{number of elements in cluster } j}{n}$.

This initialization algorithm is constructed for large databases. It is only feasible to apply the fuzzy C means method and method of moments once for small data sets. Next, reference [12] introduces the **Dirichlet Mixture**

Estimation Algorithm for estimating the parameters of a Dirichlet mixture:

1. Input k-dimensional data x_i^k, $i = 1, \ldots, n$. Also, set the number of clusters to m.
2. Use the **Initialization Algorithm**.
3. Update α_j^k by using equation (4.9), $j = 1, \ldots, m$.
4. Update $p(j)$ by using equation (4.6), $j = 1, \ldots, m$.
5. Discard component j if $p(j) < \epsilon$, go to step 3.
6. Terminate the algorithm if the convergence test is passed, else go to step 3.

Reference [13] provides a statistical method for the test of convergence when the sample size is sufficiently large. Consider the test statistic S below:

$$\begin{aligned} S = & \left(\frac{\partial}{\partial \hat{\beta}_{jl}} \Phi(x^k, \Theta, \Lambda) \quad \cdots \quad \frac{\partial}{\partial \hat{\beta}_{jk}} \Phi(x^k, \Theta, \Lambda) \right) \\ & + \begin{pmatrix} VAR(\hat{\beta}_{jl}) & \cdots & COV(\hat{\beta}_{jl}, \hat{\beta}_{jk}) \\ \vdots & \ddots & \vdots \\ COV(\hat{\beta}_{jk}, \hat{\beta}_{j1}) & \cdots & VAR(\hat{\beta}_{jk}) \end{pmatrix} \\ & \times \begin{pmatrix} \frac{\partial}{\partial \hat{\beta}_{jl}} \Phi(x^k, \Theta, \Lambda) \\ \vdots \\ \frac{\partial}{\partial \hat{\beta}_{jk}} \Phi(x^k, \Theta, \Lambda) \end{pmatrix}, \quad j = 1, \ldots, m. \end{aligned} \tag{4.12}$$

S can be shown is Chi-square distributed with k degrees of freedom. The convergence test is passed when S is smaller than $\chi_k^2(v)$ for a fixed v.

4.4 Experimental Results

We have previously reviewed the algorithm to estimate parameters of Dirichlet mixtures. Next, we will discuss three evaluation procedures to test the performance of the Dirichlet Mixture Estimation Algorithm.

A non-contextual evaluation is discussed first in [12]. Non-contextual evaluation focuses on the estimation of artificial histograms. First, consider an artificial Dirichlet mixture defined as

$$p(x_i) = \sum_{j=1}^{m} \frac{\Gamma(\alpha_{j1} + \alpha_{j2})}{\Gamma(\alpha_{j1})\Gamma(\alpha_{j2})} x_i^{\alpha_{j1}-1}(1 - x_i)^{\alpha_{j2}-1}, \tag{4.13}$$

where $i = 1, \ldots, n$. We can generate a histogram which contains n data points from this artificial Dirichlet mixture, then apply the data points to the Dirichlet Mixture Estimation Algorithm. Thus, we can obtain the estimated parameters of the Dirichlet mixture. There are three histograms generated in [12] (see Figure 4.1,4.2 and 4.3). We can evaluate the algorithm by checking the bias between the real parameters and the estimated parameters of each histogram. This results are shown in Figure 4.4 (Tables I, II, and III) which is from [12] and we see that the algorithm gives good estimates of the parameters. Note that the sample size of the first, second and third artificial histogram is 100, 255 and 255 respectively. Also, the authors in [12] found the exact number of components by taking $M = 5$ and applied the three criteria (AIC, BIC and MDL).

The second validation is a contextual evaluation. This contextual evaluation focuses on the summarization of image databases. Summarization

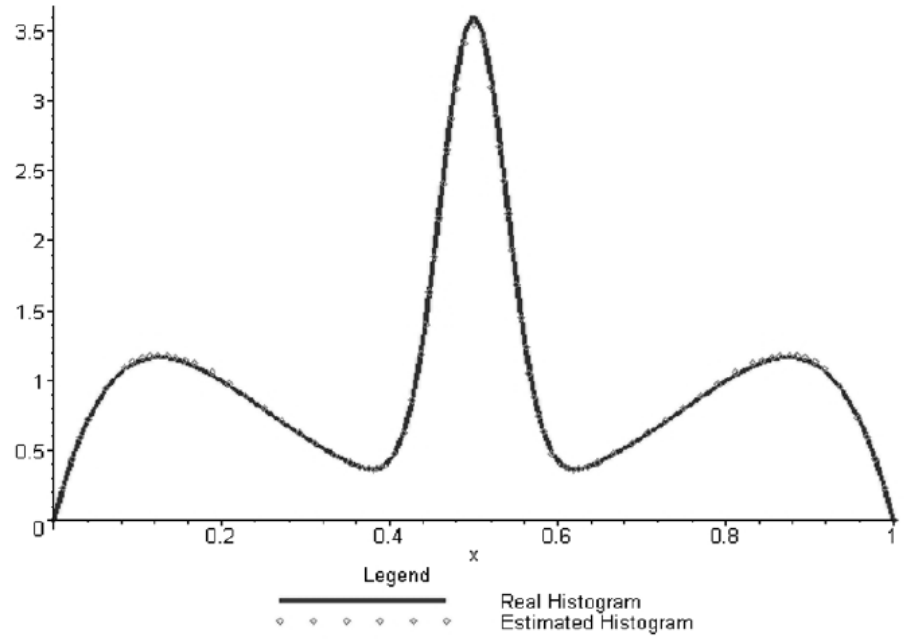


Figure 4.1: The first artificial histogram in [12]

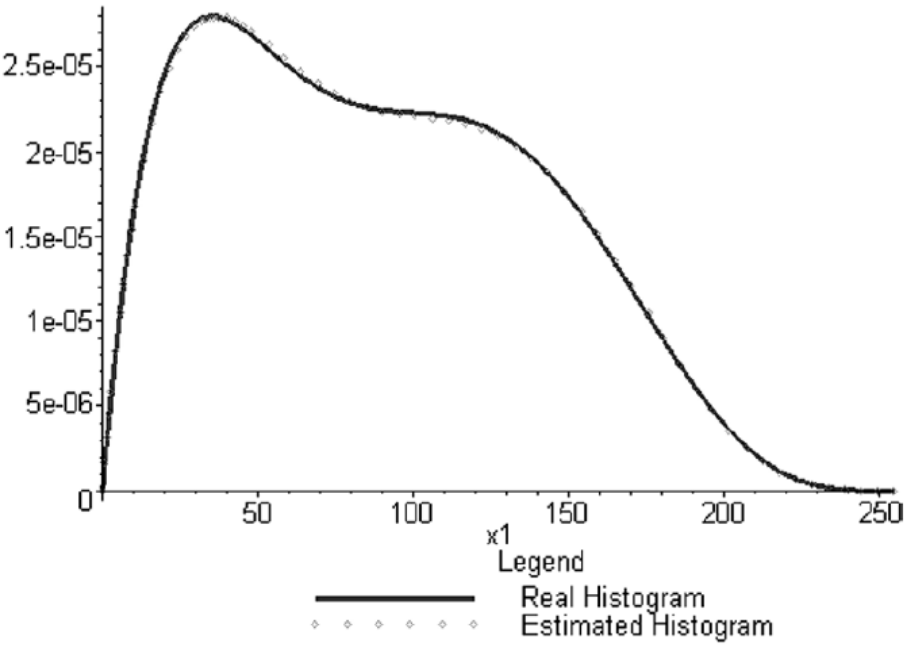


Figure 4.2: The second artificial histogram in [12]

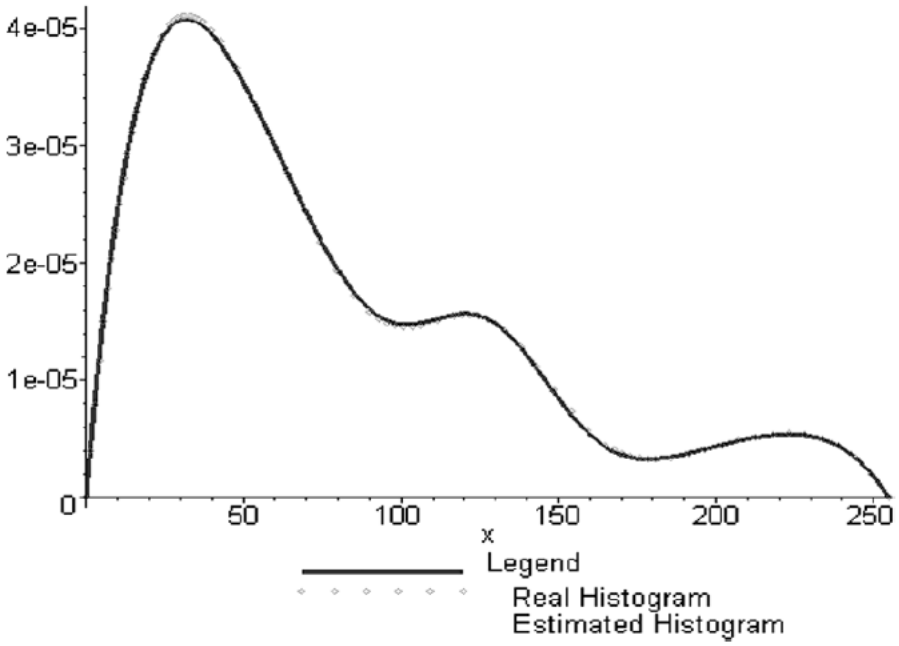


Figure 4.3: The third artificial histogram in [12]

restricts a smaller domain of the database when people search for similar images. Therefore, summarization of image databases makes the task of retrieval more efficient. Mixture decomposition can be used to find natural groupings of images and choose the most representative image to show each group. In particular, the authors in [12] extracted appropriate features from images and also partition the feature space into regions. Summarization is accomplished by identifying the homogeneous regions in the feature space. A database with 600 images (size 128×96 pixels) are used in [12]. Colors are chosen as a feature and pixels are projected on the $3D$ HSI (H = hue, S = saturation and I = intensity) space to determine the characteristic vector of each image. Thus, a $3D$ color histogram is obtained for each image. Furthermore, an $8D$ feature vector is obtained from the $3D$ color histogram

TABLE I
ESTIMATION OF THE PARAMETERS OF THE FIRST ARTIFICIAL HISTOGRAM

	Real parameters	Estimated parameters
	$P(1)$=0.33	$P(1)$=0.333
Mode 1	α_{11} = 8	α_{11} = 8.116
	α_{12} = 2	α_{12} = 2.024
	$P(2)$=0.34	$P(2)$=0.335
Mode 2	α_{21} = 80	α_{21} = 79.567
	α_{22} = 80	α_{22} =79.567
	$P(3)$=0.33	$P(3)$=0.332
Mode 3	α_{31} = 2	α_{31} = 2.024
	α_{32} = 8	α_{32} = 8.116

TABLE II
ESTIMATION OF THE PARAMETERS OF THE SECOND ARTIFICIAL HISTOGRAM

	Real parameters	Estimated parameters
	$P(1)$=0.5	$P(1)$=0.545
Mode 1	α_{11} = 2	α_{11} = 1.942
	α_{12} = 8	α_{12} = 7.141
	$P(2)$=0.5	$P(2)$=0.455
Mode 2	α_{21} = 5	α_{21} = 5.466
	α_{22} = 5	α_{22} =5.108

TABLE III
ESTIMATION OF THE PARAMETERS OF THE THIRD ARTIFICIAL HISTOGRAM

	Real parameters	Estimated parameters
	$P(1)$=0.75	$P(1)$=0.746
Mode 1	α_{11} = 2	α_{11} = 2.043
	α_{12} = 8	α_{12} = 8.307
	$P(2)$=0.15	$P(2)$=0.156
Mode 2	α_{21} = 20	α_{21} = 19.185
	α_{22} = 20	α_{22} =19.086
	$P(3)$=0.1	$P(3)$=0.098
Mode 3	α_{31} = 8	α_{31} = 8.220
	α_{32} = 2	α_{32} = 2.035

Figure 4.4: Parameters estimation results of three histograms from [12]

in [12]. The method to obtain $8D$ vectors [27] is subdividing the H, S and I axes into 2 equal intervals. Therefore, a $8D$ feature vector can represent each image.

The authors in [12] apply the algorithm and three criteria (AIC, BIC and MDL) to the feature vectors and determine the number of classes (see Figure 4.5). The number of classes is chosen by the smallest value of three criteria. There are 5 classes of 600 images, which is consistent with a human subject summarization result. A confusion matrix is created to evaluate the performance of the Dirichlet Mixture Estimation Algorithm. Cell (class i, class j) in this confusion matrix represents the number of images that are classified as class j while these images are from class i. The number of misclassified images is counted by comparing the summarization result using the Dirichlet Mixture Estimation Algorithm with the one generated by the human subject [12]. The number of misclassifications is 40 and the accuracy of this classification is $\frac{560}{600} = 93.33\%$ (see Figure 4.6).

The third validation is also a contextual evaluation. This contextual evaluation focuses on detecting human skin regions in color images. Reference [12] states that the major difference between different skin colors is intensity not color itself. Chromatic colors (pure colors in the absence of luminance) are used to represent color images. It is defined as below:

$$r_1 = \frac{r}{r+g+b}, \quad g_1 = \frac{g}{r+g+b}, \quad b_1 = \frac{b}{r+g+b}, \tag{4.14}$$

where r, g and b represents red, green and blue. Consider a training set containing more than six hundred images. Each image contains a different human skin color. There are 10867932 skin color pixels used to build a skin color model [12]. Each pixel consists of three values (r_1, g_1, b_1). Now given an image, we want to detect the skin region of the image. In order to obtain

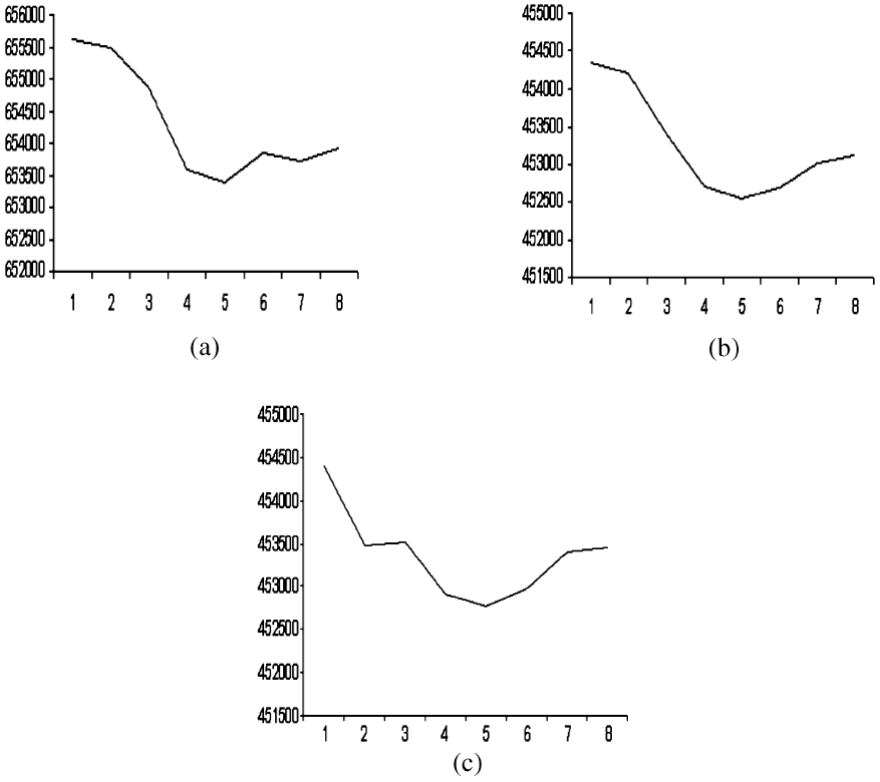


Figure 4.5: Number of classes found by the three criteria: (a) AIC, (b) MDL and (c) BIC from [12]

TABLE IV
CONFUSION MATRIX FOR IMAGE CLASSIFICATION BY A DIRICHLET MIXTURE

	Class1	Class2	Class3	Class4	Class5
Class1	101	0	10	0	0
Class2	0	120	0	13	0
Class3	0	0	108	0	0
Class4	0	6	0	104	4
Class5	0	2	0	5	127

Figure 4.6: Confusion matrix for the Dirichlet mixture from reference [12]

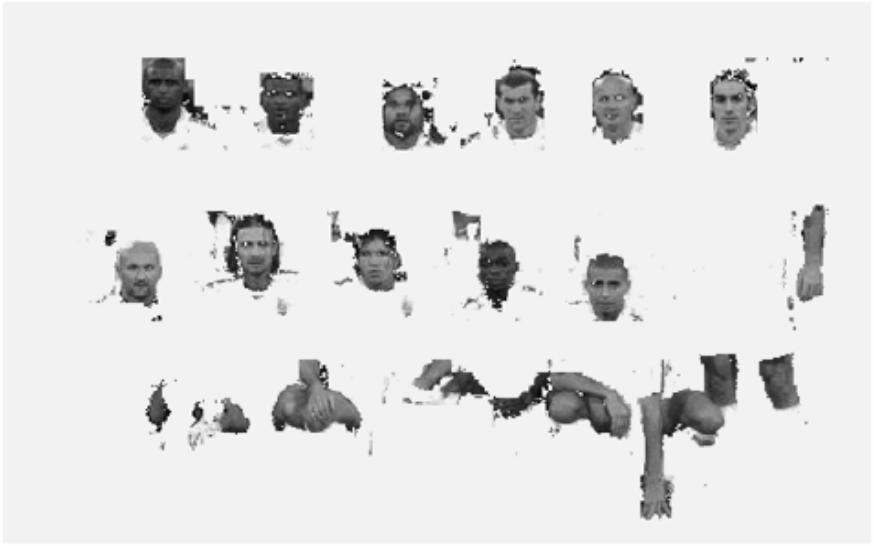
homogeneous regions, we need to segment the image and extract features. If the probability measure of a pixel is above a threshold, then that pixel is classified as a certain skin color. In this application, a region is recognized as a skin area if more than 75% of pixels in that region are classified as skin color. Results of skin detection by using a mixture of Dirichlet distributions and a mixture of Gaussian distributions are shown in [12] (see figure 4.7,4.8 and 4.9). To compare which mixture model is more accurate in detecting skin area, we notice that figure 4.9 (skin area extracted using a Dirichlet mixture) contains less non-skin area than figure 4.8 (skin area extracted using a Gaussian mixture).

(a)

Figure 4.7: Original image in [12]

(b)

Figure 4.8: Skin area extracted using a Gaussian mixture in [12]



(c)

Figure 4.9: Skin area extracted using a Dirichlet mixture in [12]

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this report, we have presented the definition and properties of the Dirichlet distribution. These properties include the derivation of information-theoretic quantities, such as differential entropy, divergence and mutual information, for the Dirichlet distribution. The stick-breaking approach and the Pólya urn method are discussed for generating random variables with a Dirichlet distribution. The notion of exchangeability is introduced and the De-Finetti's theorem, stating that an exchangeable binary sequence is a mixture of i.i.d Bernoulli sequences is examined. We have also discussed the Pólya Urn model with two different color balls, a fixed number of color balls and infinitely many color balls. Furthermore, we have shown that the Dirichlet distribution and the generalized Dirichlet distribution can be used as a prior distribution due to its conjugacy property. Moreover, we have discussed one application of the Dirichlet distribution based on [12]. We have

described an unsupervised learning algorithm for a Dirichlet mixture model with multivariate data. The Initialization and Dirichlet Mixture Estimation Algorithms of [12] are reviewed for estimating the parameters of a Dirichlet mixture. Three experimental results of [12] show that the Dirichlet mixture excels at modeling data.

5.2 Future Work

Future work may consider the use of the Bayesian approach to estimate parameters of a Dirichlet mixture model as seen in Chapter 4. Also, we can consider the use of some Dirichlet related distributions to model multivariate data, such as the nested Dirichlet distribution and the Dirichlet-Multinomial distribution.

Bibliography

[1] Ken Aho, DeWayne Derryberry, and Teri Peterson. Model selection for ecologists: the worldviews of AIC and BIC. *Ecology*, 95(3):631–636, 2014.

[2] Haider Al-Lawati and Fady Alajaji. On decoding binary perfect and quasi-perfect codes over markov noise channels. *IEEE Transactions on Communications*, 57(4):873–878, 2009.

[3] Fady Alajaji and Tom Fuja. A communication channel modeled on contagion. *IEEE Transactions on Information Theory*, 40(6):2035–2041, 1994.

[4] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

[5] Ghady Azar and Fady Alajaji. On the equivalence between maximum likelihood and minimum distance decoding for binary contagion and queue-based channels with memory. *IEEE Transactions on Communications*, 63(1):1–10, 2015.

[6] Narayanaswamy Balakrishnan and Valery B Nevzorov. *A primer on statistical distributions*, chapter 27. John Wiley & Sons, 2004.

[7] Amit Banerjee, Philippe Burlina, and Fady Alajaji. Image segmentation and labeling using the polya urn model. *IEEE Transactions on Image Processing*, 8(9):1243–1253, 1999.

[8] Arindam Banerjee and Sugato Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *SDM*, volume 7, pages 437–442. SIAM, 2007.

[9] José M Bernardo and Adrian FM Smith. *Bayesian theory*, chapter 4. John Wiley & Sons, 1994.

[10] David Blackwell and James B MacQueen. Ferguson distributions via pólya urn schemes. *The Annals of Statistics*, pages 353–355, 1973.

[11] Nizar Bouguila and Djemel Ziou. A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized Dirichlet mixture. *IEEE Transactions on Image Processing*, 15(9):2657–2668, 2006.

[12] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Unsupervised learning of a finite mixture model based on the Dirichlet distribution and its application. *IEEE Transactions on Image Processing*, 13(11):1533–1543, 2004.

[13] SC Choi and R Wette. Maximum likelihood estimation of the parameters of the gamma distribution and their bias. *Technometrics*, 11(4):683–690, 1969.

[14] Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[15] Thomas M Cover and Joy A Thomas. *Elements of information theory*, chapter 2, 17. John Wiley & Sons, 2012.

[16] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[17] Luc Devroye. Sample-based non-uniform random variate generation. In *Proceedings of the 18th Conference on Winter Simulation*. ACM, 1986.

[18] Joseph C Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. 1973.

[19] Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[20] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1, chapter 14. Springer series in statistics Springer, Berlin, 2001.

[21] Bela A Frigyik, Amol Kapila, and Maya R Gupta. Introduction to the Dirichlet distribution and related processes. Technical report, UWEETR-2010-0006, 2010.

[22] Zoubin Ghahramani. Unsupervised learning. In *Advanced lectures on machine learning*, pages 72–112. Springer, 2004.

[23] Gerald Joseph Goodhardt, Andrew SC Ehrenberg, and Christopher Chatfield. The Dirichlet: A comprehensive model of buying behaviour. *Journal of the Royal Statistical Society. Series A (General)*, pages 621–655, 1984.

[24] Mark H Hansen and Bin Yu. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

[25] Robert V Hogg and Allen T Craig. *Introduction to mathematical statisties 7th ed*, chapter 3. 1970.

[26] Mohamed Maher Ben Ismail and Hichem Frigui. Unsupervised clustering and feature weighting based on generalized Dirichlet mixture modeling. *Information Sciences*, 274:35–54, 2014.

[27] Mohammed Lamine Kherfi, Djemel Ziou, and Alan Bernardi. Combining positive and negative examples in relevance feedback for content-based image retrieval. *Journal of Visual Communication and Image Representation*, 14(4):428–457, 2003.

[28] Kenneth Lange. Applications of the Dirichlet distribution to forensic match probabilities. *Genetica*, 96(1-2):107–117, 1995.

[29] Tao Li, Sheng Ma, and Mitsunori Ogihara. Entropy-based criterion in categorical clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 68. ACM, 2004.

[30] Nicholas T Longford. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827, 1987.

[31] Hosam Mahmoud. *Pólya urn models*, chapter 1. CRC press, 2008.

[32] Tom Minka. Bayesian inference, entropy, and the multinomial distribution. 2003.

[33] Sayan Mukherjee. Exchangeability. Fall 2009.

[34] Kai Wang Ng, Guo-Liang Tian, and Man-Lai Tang. *Dirichlet and related distributions: Theory, methods and applications*, volume 888, chapter 1, 2. John Wiley & Sons, 2011.

[35] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[36] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *ArXiv Preprint ArXiv:0911.4863*, 2009.

[37] Brad Null. Modeling baseball player ability with a nested Dirichlet distribution. *Journal of Quantitative Analysis in Sports*, 5(2), 2009.

[38] Robin Pemantle et al. A survey of random processes with reinforcement. *Probab. Surv*, 4(0):1–79, 2007.

[39] Mark J Schervish. *Theory of statistics*, chapter 1. Springer Science & Business Media, 2012.

[40] Samuel S Wilks. Mathematical statistics. 1962. *New York, John Wileyand Sons.*

[41] Tzu-Tsung Wong. Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, 97(2):165–181, 1998.

[42] Libo Zhong, Fady Alajaji, and Glen Takahara. A binary communication channel with memory based on a finite queue. *IEEE Transactions on Information Theory*, 53(8):2815–2840, 2007.

[43] Zoran Zivkovic and Ferdinand van der Heijden. Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):651–656, 2004.