View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Concordia University Research Repository

Distributions based Regression Techniques for Compositional Data

Divya Ankam

A Thesis

in

The Concordia Institute

for

Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

February 2019

© Divya Ankam, 2019

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Divya Ankam**

Entitled: **Distributions based Regression Techniques for Compositional Data**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Chun Wang

_____ External Examiner
Dr. Govind Gopakumar

_____ Examiner
Dr. Fereshteh Mafakheri

_____ Supervisor
Dr. Nizar Bouguila

Approved by _____
Abdessamad Ben Hamza, Chair
Department of Information Systems Engineering

_____ 2019

_____
Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Distributions based Regression Techniques for Compositional Data

Divya Ankam

A systematic study of regression methods for compositional data, which are unique and rare are explored in this thesis. We start with the basic machine learning concept of regression. We use regression equations to solve a classification problem. With partial least squares discriminant analysis (PLS-DA), we follow regression algorithms and solve classification problems, like spam filtering and intrusion detection. After getting the basic understanding of how regression works, we move on to more complex algorithms of distributions based regression. We explore the uni-dimensional case of distributions, applied to regression, the beta-regression. This gives us an understanding of how, when the data to be predicted, or the outcome, is assumed to be of beta distribution, a prediction can be made with regression equations. To further enhance our understanding, we look into Dirichlet distribution, which is for a multi-dimensional case. Unlike traditional regression, here we are predicting a compositional outcome. Two novel regression approaches based on distributions are proposed for compositional data, namely generalized Dirichlet regression and Beta-Liouville regression. They are extensions of Beta regression in a multi-dimensional scenario, similar to Dirichlet regression. The models are learned by maximum likelihood estimation algorithm using Newton-Raphson approach. The performance comparison between the proposed models and other popular solutions is given and both synthetic and real data sets extracted from challenging applications such as market share analysis using Google-Trends and occupancy estimation in smart buildings are evaluated to show the merits of the proposed approaches. Our work will act as a tool for product based companies to estimate how their investments in advertising have yielded results in the market shares. Google-Trends gives an estimate of the popularity of a company, which reflects the effect of advertisements. This thesis bridges the gap between open source data from Google-Trends and market shares.

Acknowledgments

The journey towards the completion of my thesis was packed with excitement, learning, friendships, fights, discipline, sacrifices, struggles and joys. I learnt life is like running a marathon and not sprints.

My sincere gratitude goes towards my thesis supervisor, Dr. Nizar Bouguila. He polished me into the researcher and writer I am today. He believed in me when I was drowning in self doubt. His patience, persistence and most importantly, freedom of thought, have made me find myself achieving my dreams in Canada, a land far away from my home, India.

I thank my professors, staff and colleagues of Concordia Institute for Information Systems Engineering for their support.

Friends are family when you are on foreign lands! I thank all my lab-mates. Especially I would like to thank Muhammad Azam, Nuha Zamzami, Narges Manouchehri, Kamal Maanicshah and Soodeh Akbari.

My daughter Disha Spurthi, my husband Renu Babu, his parents, my parents, my grandparents and everyone in our family have always supported me and stood by my side to make this achievement possible. I am heavily indebted to them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In machine learning, we can say there are three main classical approaches, regression, classification and clustering. A systematic study of regression methods for compositional data, which are unique and rare are explored in this thesis. We start with the basic machine learning concept of regression. We use regression equations to solve a classification problem. With partial least squares discriminant analysis (PLS-DA), we follow regression algorithms and solve classification problems, like spam filtering and intrusion detection. After getting the basic understanding of how regression works, we move on to more complex algorithms of distributions based regression. We explore the uni-dimensional case of distributions, applied to regression, the beta-regression. This gives us an understanding of how, when the data to be predicted, or the outcome, is assumed to be of beta distribution, a prediction can be made with regression equations. To further enhance our understanding, we look into Dirichlet distribution, which is for a multi-dimensional case. Unlike traditional regression, here we are predicting a compositional outcome. For example, with one input, the depth of Arctic lake, we predict the soil composition of the Arctic lake sediments. The sediments are composed of a portion of sand, silt and clay. There are 3 different outputs to be predicted in this scenario. Classical methods will not be able to answer these questions, hence, distributions based regression come into play. Interestingly, in this research arena, the research on multidimensional distributions based regression stopped at Dirichlet regression. We take the study further into generalized Dirichlet regression and Beta-Liouville regression. To prove our efficiency, we have chosen a couple of real world examples and also an interesting application of how we can predict share

market dynamics with the help of Google-trends.

Market share means the fraction of the market held by a company for the entire sales happening for that particular product. For example, if for this month, a hundred mobile phones were sold in a city, and if Apple company sold 80 phones out of them, then we can say Apple holds 80% of mobile phone market shares in that city for that particular month. At present, when a company wants to know how have its investments in advertisements have been fruitful in giving it a raise in the market shares, it would have to look into the balance sheet of investments versus market shares. There is no way of finding how much its competitors have invested in advertisements. To tackle this problem, we have looked into Google-Trends. Googe-Trends is a golden trove of user web searches. Here we can know how many time a company/brand name was googled for, and also how it fared in comparison to its competitors. When a company has advertised its product, people are attracted to look it up more in the internet, if the advertisement has reached them. Internet marketing is the most common type of marketing in today's era. more people are using the internet compared to news-papers, roadside Ad-posts or pamphlets. Internet advertisements have the most public reach. Utilizing the open source data available to us on Google-trends, we can estimate the advertising investments of a company as well as its competitors. Using this method we propose in our thesis, the companies are benefited with the most useful information of how popular they are in comparison to their competitors, and they could also compare themselves with their past performance. So to summarize this link, the company invests in advertisements, users are attracted to the ads and look up the company name term in the internet, this is recorded by Google-trends, our works gives a tool to compare this data with the market shares performance of the company in question. The regression methods we propose are novel and also this approach we explained are unique to our thesis.

1.2 Introduction

Compositional data are naturally generated by many applications from different domains. Examples of classic domains include analytical chemistry, geology, petrology, sedimentology, and metabolomics [1]. Moreover, this kind of data has attracted recently attention in computer vision [2, 3], pattern recognition [4, 5, 6, 7], and machine learning applications [8, 9, 10, 11, 12, 13]. Compositional data have a constant sum property while all components are positive real numbers.

When the sum is equal to one, they are called proportional data [9]. In the study of ground-water samples, Hydrogeology, scientists are interested in describing and understanding the composition of water samples. It is more meaningful to express fluoride content of a sample in parts per million or percentage. It is not of potential interest to know that a sample contains 1mg of fluoride, rather 5mg/L of fluoride (data on absolute frequencies) indicates toxic levels. In the analysis of images [14, 15], it is common to represent a given image as a histogram representing the proportions of grey levels [16, 17, 18]. In data mining, textual documents are often represented as vectors representing the proportions of some keywords [19].

Mathematically compositional data are represented in a standard simplex of the sample space given by,

$$S^D = \{x = [x_1, x_2, \ldots, x_D] \in \mathbb{R}^D\} \tag{1}$$

where $x_i > 0, i = 1, \ldots, D$ and $\sum_{i=1}^{D} x_i = k$; x is a D-dimensional vector of features representing a given object (e.g. document, image, video, etc.) and k is a constant. Due to the compositional constraint the data points cannot be analyzed on the Euclidian space, simplex sample space is a better projection of the data [20, 21, 22]. To analyze compositional data in Euclidian space and for further analysis with known classification algorithms, it is necessary to convert those using transformations to such co-ordinate system [23]. For instance, Log-ratio approach was proposed to transform data from simplex to general coordinates, without the loss of relative scale property. Thereafter, additive, centered and isometric log-ratio transformations are most widely used . Later a data-based power transformation (α- transformation) for compositional data was proposed by [24]. It was a compromise between the raw data analysis and log-ratio analysis.

1.3 Contributions

The contributions of this thesis are as follows:

- **α-tranformation with PLS-DA and Applications:** In compositional data, the relative proportions of the components contain important relevant information. In such case, Euclidian distance fails to capture variation when considered within data science models and approaches such as partial least squares discriminant analysis (PLS-DA). Indeed, the Euclidean distance assumes implicitly that the data is normally distributed which is not the case of compositional vectors. Aitchison transformation has been considered as a standard in compositional data

analysis. In this chapter, we consider two other transformation methods, Isometric log ratio (ILR) transformation and data-based power (alpha) transformation, before feeding the data to PLS-DA algorithm for classification [25]. In order to investigate the merits of both methods, we apply them in two challenging applications namely spam filtering and intrusion detection. This work has been published in [26].

- **Generalized Dirichlet Regression applied to Market-Shares:** We explore the idea that market-shares of any given company have a linear relationship with the number of times the company/product is searched for on the internet. This relationship is critical in deducing whether the funds spent by a firm on advertisements have been fruitful in increasing the market-share of the company. To deduce the expenditure on advertisement, we consider google-trends as a replacement resource. We propose a novel regression algorithm, generalized Dirichlet regression, to solve the resulting problem with information from three different information-technology fields: internet browsers, mobile phones and social networks. Our algorithm is compared to Dirichlet regression and ordinary-least-squares regression with compositional transformations. Our results show both the relationship between market-shares and google-trends, and the efficiency of generalized Dirichlet regression model.

- **Beta-Liouville Regression and Applications:** We propose a novel regression algorithm, Beta-Liouville regression, to solve regression of compositional data, where the prediction is multi-dimensional and sums to unity. Applications include market-share data mining in relation to its score in Google-trends and smart building occupancy estimation. Occupant behaviour in buildings gives useful insights on the required levels of air conditioning, lighting and even initiating help during emergency. Sensors to estimate the occupancy of a smart building include microphone, door/window positions, motion detection, power consumption. The Beta-Liouville regression algorithm is compared to ordinary least squares regression with compositional transformations and Dirichlet regression.

1.4 Thesis Overview

The rest of this thesis is organized as follows:

- Chapter 2 introduces different transformations of compositional data, implementation of Partial Least Squares Discriminant Analysis, applied to spam filtering and intrusion detection problems.
- Chapter 3 is devoted to Generalized Dirichlet Regression and its applications to market share analysis.
- Chapter 4 introduces Beta-Liouville Regression, applied to market shares and occupancy estimation of smart buildings.
- Chapter 5 concludes and summarizes the thesis and points out future research directions.

Chapter 2

Compositional Data Analysis with PLS-DA

2.1 Introduction

In this chapter, we focus on Partial least squares discriminant analysis (PLS-DA) when applied in the case of compositional data. PLS-DA is devoted to a unique regression problem, where the response is formed by categorical variables. Aim of the discriminant function is to decide which group a sample belongs to using its composition profile. This technique can also be considered as a compromise between linear discriminant analysis (LDA) and DA on the significant principal components of the predictor variables [27]. A statistical explanation regarding the relation of PLS to canonical correlation analysis (CCA), which in turn is related to LDA is given in [28]. PLS-DA has been generally applied for Gaussian data. In this chapter, we propose some approaches to handle compositional data. We apply the resulting models to two challenging problems namely spam filtering and intrusion detection. This work has been accepted as a conference paper [26].

2.2 Partial Least Square Discriminant Analysis (PLS-DA)

PLS-DA is a special kind of regression analysis used for classification problems. The main goal is to find a line/hyperplane that acts as partition between classes in a dataset. This is performed by the partial least squares method. PLS-DA is chosen over Principal Component Analysis (PCA) when dimensionality reduction is necessary along with classification [28]. PLS-DA works similar

to a supervised version of PCA. In case of a PCA algorithm, the first principal component along the first eigenvector is calculated by minimizing the projection error, in turn maximizing the variance of projected data [29]. The algorithm iteratively projects all the points to a subspace orthogonal to the last principal component and then repeats the process on the projected points. Further an orthonormal basis of eigenvectors and principal components is formed. PLS-DA aims at finding uncorrelated linear transformations, (latent components) of the original predictor variables, which have high covariance with the response variables. Based on these latent components, PLS predicts response variables Y and reconstructs original matrix X at the same time. In case of PLS-DA, its principal components are linear combinations of features [30]; the number of these components gives the dimension of the transformed space. As a standard, the components are orthogonal to each other. The iterative process computes the transformation vectors also known as loading vectors, which give the importance of each feature in that component. There are several algorithms for Partial least squares (PLS) and its enhancements for Discriminant analysis. A comparison of nine PLS1 algorithms is given in [31].

Consider a data set with n objects. Each object is described by D features. Matrix X of dimension n times D contains explanatory variables. The response variables are stored in Y (denoted as c here), an n times q dimensional matrix, corresponding to n classes. Y consists of binary variables. For example, in a two-class classification case, if the data point corresponds to class-I then Y vector is (1, 0), if the point belongs to class-II then Y vector is (0, 1). Before feeding to the regression model, Y is mean centered as is the usual case. Aitchison geometric mean centering is performed on X due to compositional constraint. X transformations are explained in section III. The PLS can be formalized as following [32]. PLS-DA is derived from PLS regression and involves forming a regression model between X and c. PLS-DA is similar to any other linear decision function but is often described algorithmically rather than statistically. Fundamental PLS-DA equations:

$$\begin{aligned} X &= T\,P + E \\ c &= T\,q + f \end{aligned} \tag{2}$$

T is the common score matrix, E and f are residuals of X and c, respectively. p and q are loadings of X and c, respectively. By the decomposition of X and c, response values are decided by the latent

variables and not by *X* alone. This makes it a more reliable model because the latent variables coincide with the underlying structure of original data. The major point of PLS-DA is the construction of components by projecting *X* on weights. PLS criterion is used to sequentially maximize the covariance between response variables and latent components. Following is the algorithm and model building of PLS-DA.

2.2.1 Outer modeling

(1) Calculate the PLS weight vector w, as:

$$w = X'c \tag{3}$$

(2) Calculate the X scores, given by:

$$t = \frac{Xw}{\sqrt{\sum w^2}} \tag{4}$$

In PLS, line of best fit is calculated, in which the sum of squares of *X* residuals is minimized. Regards to principal component, loadings are the angle cosines of the direction vector. Scores are the projections of the sample points on the principal component direction [33].

(3) Calculate C loading by

$$q = \frac{c't}{\sum t^2} \tag{5}$$

(4) Calculate *Y* scores, given by

$$u = c\,q \tag{6}$$

2.2.2 Residual Deflation

(1) Regression coefficient vector is estimated as

$$b = \frac{u't}{\sum t^2} \tag{7}$$

(2) Calculate *X* loadings by

$$p = \frac{t'X}{\sum t^2} \tag{8}$$

(3) By subtracting the effect of the new PLS component from the data matrix, we obtain a residual data matrix

$$^{resid}X = X - tp \tag{9}$$

(4) Calculate the Residual value of c

$$^{resid}c = c - tq \tag{10}$$

Replace both X and c by their residuals and return to step 1 if further components are required.

2.2.3 Prediction

Once, the model has been built, prediction of class can be done for original or new data. The regression coefficients Γ of dimension D times q is given by

$$B = pbq' \tag{11}$$

The PLS regression equation, when X is mean centred is shown as

$$Y = XB + E \tag{12}$$

For compositional data, where X centring is carried out in Aitchison geometry as explained in further sections, the linear regression takes the form

$$Y = \Gamma B + E \tag{13}$$

where the regression coefficient is given by

$$\Gamma = pbq' \tag{14}$$

Training data are used to model PLS-DA and (13) is generated. Test data is substituted in (13) to predict class of testing data. The rule of classification depends on the value of testing set predictor class chosen.

2.3 Data Transformations

Statistical approaches, based on Gaussian assumption, cannot be directly implemented on compositional data due to constant sum constraint and co-relation among features, as explained in the introduction. The proportions of the species are more pertinent than the number of instances. Evidently the scale of proportions is not absolute but relative. In such a case, standard mean values and variances are no longer valid, and a suitable transformation is used to convert into a new scale, which can be assumed near to absolute [34]. We will consider three transformations in this chapter and they are discussed in the following subsections.

2.3.1 CLR - Centered Log Ratio Transform

The centered log ratio transform [23] is defined as:

$$y = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)' \tag{15}$$

CLR can also be represented as follows

$$y = \{y_i\}_{i=1,\ldots,D} = \left\{ \ln \frac{x_i}{g(x)} \right\}_{i=1,\ldots,D} \tag{16}$$

$$g(x) = \Big(\prod_{j=1}^{D} x_j \Big)^{1/D} \tag{17}$$

Here *g(x)* is the geometric mean of the composition. Resulting in *D* variables, each representing one component of the original compositional part, it is easily interpretable as to what is the contribution of each part individually. CLR has drawbacks – it generates a singular matrix, which is sum of resulting parts is zero, sub-compositional incoherence is also an issue. Most of the present statistical analysis methods cannot be performed on singular data. This can be overcome by ILR transformation.

2.3.2 ILR - Isometric Log Ratio transform

The isometric log ratio transform is defined as

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, i = 1, \ldots, D-1 \tag{18}$$

This results in *D*-1 coordinates in a chosen orthonormal basis. It represents only the ratios between the components. ILR can also be represented in terms of CLR as

$$z = Hy \tag{19}$$

where *H* is the Helmert sub-matrix, obtained by removing the first row of Helmert matrix [35]. It is the most sought after orthonormal basis. ILR is applied on data matrix *X*, subsequent ILR co-ordinates *Z* are fed into the PLS-DA algorithm in section II.A. The regression coefficient matrix thus obtained, using equation (3) is Γ, has $(D-1)$ times n dimensions. When substituted in equation (3), the predictor variables *Y* are obtained.

2.3.3 Data-based power transformation

The data-based power transform, also known as α-transform is defined as

$$u = \{u_i\}_{i=1,\ldots,D} = \left\{ \frac{x_i^{\alpha}}{\sum_{j=1}^{D} x_j^{\alpha}} \right\}_{i=1,\ldots,D} \tag{20}$$

$$z = \frac{1}{\alpha}(Du - 1)\, H^T \tag{21}$$

Value of α lies between 0 and 1, chosen based on the data set itself [36]. As $\alpha \to 0$ the equation converges to ILR transform. As α=1, it ignores the compositional constraint and treats the data as if it were Euclidian. It is advantageous over ILR as it can be applied to datasets containing elements with zero value, whereas ILR is not defined for them.

2.4 Experimental Results

In this section we present our experimental results through two real-world challenging applications. The first one concerns spam filtering and the second one deals with intrusion detection [37].

The below procedure is followed to convert the data generated by both applications to compositional form. Features where relative frequencies with which the components occur is measurable are selected. The components selected should be fractions of a whole. To obtain a D-part composition, we divide the variables of the dataset by the sum of all variables.

$$x_c = \{x_i\}_{i=1,\ldots,D} = \left\{ \frac{x_i}{\sum_{j=1}^{D} x_j} \right\}_{i=1,\ldots,D} \tag{22}$$

For each of the transformations, the steps can be summarized as follows

(1) Split data into training and testing sets.

(2) Apply a transformation as in Section 2.3.

(3) Feed testing data to PLS-DA algorithm as in Section 2.2 to obtain PLS equation with regression coefficient as in Equation (14).

(4) Apply transformation to testing set, substitute in Equation (14) to get predictor variable Y.

(5) Use a classification decision rule to predict class.

The simplest rule of classification decision is based on the value of Y, if it is positive, assign value 1, and 0 for negative values. This is not a hard and fast rule, it may not be the most appropriate decision rule for all data sets. Further discussion on this is given in [32].

2.4.1 Spam Filtering

Spam, also called unsolicited bulk e-mail or junk email, is an email that is sent to a gathering of beneficiaries who have not requested for it. Spam can be also a serious threat. IBM Security today announced results from 2017 IBM X-Force Threat Intelligence Index. It claims that malicious attachments in a spam email are a primary delivery method of ransomware. There is 400% increase in spam every year. Ransomware made up 85% of those malicious attachments in 2016. It is important to have an email spam filter [38] for a company to save employees from frustration of an overloaded mailbox of unimportant and occasionally dangerous emails.

Several approaches have been proposed in the past for Spam filtering. Among these approaches, machine learning techniques have received particular attention. In this subsection, we apply our framework for the problem of spam filtering. We have worked on two datasets to compare the

Table 2.1: Confusion matrix for spam classification on Spambase using ILR + PLS-DA

	Normal	Attacks
Normal	884	22
Attacks	287	619

Table 2.2: Confusion matrix for spam classification on Spambase using α + PLS-DA

	Normal	Attacks
Normal	889	17
Attacks	254	652

methodology proposed. The considered datsets have been previously used for spam filtering in [33, 39]

a) **Spambase** dataset[1] used in this chapter has been obtained from UCI machine learning repository, created by Hewlett-Packard Labs. This dataset has 4601 instances and 58 attributes (57 continuous input attributes and 1 nominal class label target attribute). The number of spam mails is 1813 (39.4%) and non-spam are 2788 (60.6%).

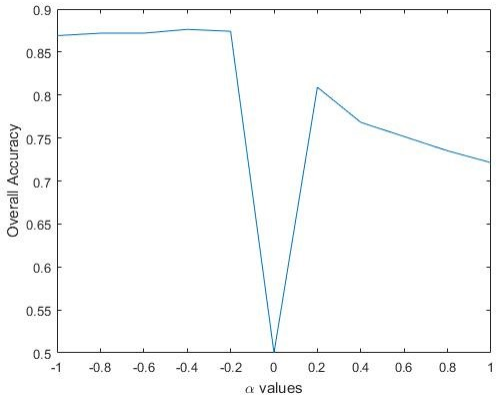
As the study presented in this chapter is for compositional data. It has to be noted that the relative ratio among features of compositional data is of key importance. In that context, it would be appropriate to consider only those features which are of the same category, which in this case is percentage of characters in the e-mail that match WORD/CHAR. There are 54 such features. The dataset has been reduced to 3626 instances to have a balanced case for optimization of PLS-DA algorithm. The feature vectors are normalized, transformed and fed to PLS-DA algorithm as in section II. Both ILR and α transformations were applied. The corresponding confusion matrices are given in Tables 2.1 2.2. The value of α is varied from 0 to 1 in steps of 0.1. The value with

[1] https://archive.ics.uci.edu/ml/datasets/spambase

Table 2.3: Test results for ILR, alpha transformations and PLS-DA on Spam datasets

Datasets	Learning Algorithm	Results			
		Overall Accuracy	Precision	Recall	False Alarm Rate
HP spambase	ILR + PLS-DA	0.8295	0.9757	0.7549	0.0343
	Alpha(0.5)+PLS-DA	0.8504	0.9812	0.7778	0.0254
Ling- spam	ILR + PLS-DA	0.9308	0.9154	0.9444	0.0821
	Alpha(0.5)+PLS-DA	0.9385	0.9308	0.9453	0.0682

Figure 2.1: Effect of $\boldsymbol{\alpha}$ on accuracy: Spambase dataset.



best result is presented in Table 2.2. Variation of accuracy with the value of α is shown in Fig 2.1. The accuracy is minimum (0.5) when α is zero, which implies the data is considered to be log ratio transformed. The accuracy of ILR + PLS-DA method is (82.95%), and for Alpha + PLS-DA is 85.04%. Though both methods report good performance, α transformation is a better choice.

b) $\mathbf{Ling-Spam}$ corpus[2] has been used to perform spam filtering. It consists of 2412 (82.4%) linguist messages and 481 (16.6%) spam messages. Ling spam corpus is a collection of emails. The data set has been split into training set of 702 emails and testing set of 260 emails, equal division of spam and legitimate emails was done to ensure a balanced case. The text data were prepared by removal of stop words and lemmatization [40]. A word dictionary created of 50 most common words. Feature extraction was done by creating word count vector with frequency of occurrence of dictionary words. The accuracy of ILR + PLS-DA method is 93.08%, and for Alpha + PLS-DA is 93.85%. There is almost similar performance with both methods. Variation of accuracy with the value of α is shown in Fig 2.2. Besides overall accuracy, we evaluate the different transformations in terms of their precision, recall and false alarm rate. Comparison of performance of both methods on the 2 spam datasets is given in 2.3. It has to be noted that α transformation outperforms ILR transformation with respect to all the evaluation measures.

Table 2.4: Confusion matrix for spam classification on Ling-spam using ILR + PLS-DA

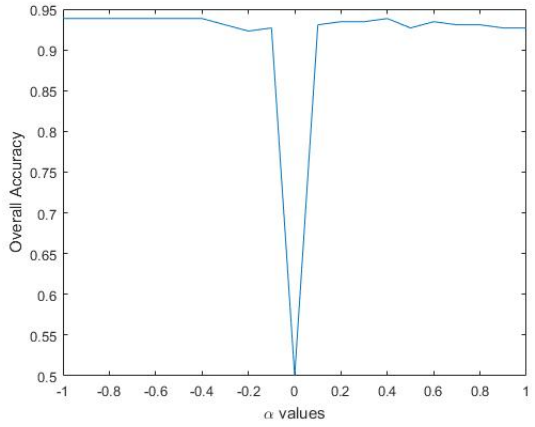
	Normal	Attacks
Normal	119	11
Attacks	7	123

[2]http://www.csmining.org/index.php/ling-spam-datasets.html

Table 2.5: Confusion matrix for spam classification on Ling-spam using α + PLS-DA

	Normal	Attacks
Normal	121	9
Attacks	7	123

Figure 2.2: Effect of α on accuracy: Ling-spam dataset.



2.4.2 Intrusion Detection

Rapid use of internet and technology has been the rule of the day. "Annual global IP traffic will reach 3.3 ZB (ZB; 1000 Exabyte [EB]) by 2021." Says Cisco executive summary[3]. Safeguarding the internet from intruders is a necessity. Intrusion detection systems (IDSs) are software or hardware systems that automate the process of monitoring and analyzing the computer system or network events for signs of security problems [41]. We have worked with the NSL-KDD dataset[4] in this chapter. The full NSL-KDD train set including labels is in the file KDDTrain+.TXT and the complete test set is from KDDTest+.TXT. Features of each network connection vector have been mentioned in detail in [42]. There are 41 features in the dataset, of which 3 are nominal, they have been enumerated. All the attacks have been grouped together. Test set contains 19,420 instances and train set has 1,17,260 instances. Dataset has been equally divided, to attain a balanced case based on normal and attack classes for optimization of PLS-DA algorithm. Tables 2.7 2.8 show the confusion matrices of the 2 different transformations applied before PLS-DA. ILR transform resulted in a poor accuracy of 64.87%. However, α transform proved to be better with 79.74% of

[3]https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[4]http://nsl.cs.unb.ca/NSL-KDD

Table 2.6: Test results for ILR, alpha transformations and PLS-DA on NSL-KDD dataset

Datasets	Learning Algorithm	Results			
		Overall Accuracy	Precision	Recall	False Alarm Rate
NSL-KDD	ILR + PLS-DA	0.6487	0.3147	0.9479	0.4108
	Alpha(0.5)+PLS-DA	0.7974	0.8371	0.7756	0.1770

emails classified correctly. As the dataset contains zero values, α transform is not defined for α = 0, hence a break in the graph is observed in Figure 2.3. Table 2.6 gives a summary of evaluation metrics of both methods on NSL-KDD dataset. Though ILR transform has a bad accuracy, its recall rate, also known as sensitivity is exceptional at 94.79% compared to 77.56% for α transform. However, keeping in mind the false rate alarm is very high for ILR, the best method would be to use α transform in conjunction with PLS-DA.

Table 2.7: Confusion matrix for NSL-KDD dataset using ILR + PLS-DA

	Normal	Attacks
Normal	3056	6654
Attacks	168	9542

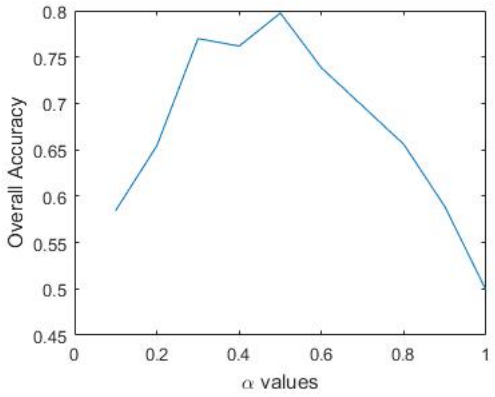
Table 2.8: Confusion matrix for NSL-KDD dataset using $\boldsymbol{\alpha}$ + PLS-DA

	Normal	Attacks
Normal	8128	1582
Attacks	2352	7358

2.4.3 Conclusions

In this chapter, we described the use of various compositional transformation techniques within PLS-DA. This is a unique combination that has not been explored before. Alpha transformation seems to perform better than ILR in all these applications. The flexibility of α value helps maximize the overall performance as shown by real-world challenging security applications namely spam classification and intrusion detection. Future work could be devoted to other applications from other domains where compositional data are naturally generated. Potential Future research could be devoted also to unbalanced compositional data that we may encounter in real applications [1].

Figure 2.3: Effect of α on accuracy: NSL-KDD dataset.



Chapter 3

Generalized Dirichlet Regression applied to Market-Shares

3.1 Introduction

Our aim is to predict the change in market-share [43] [44] composition with respect to share-of-voice on social media. We assume it is directly proportional to the investment in marketing. We are making a strong assumption that the google trends are a result of the user's search which were guided by advertisements and people talking about the company/product. We can thus deduce it to be directly proportional to the money the company spends on advertising the product [45]. The insider information on the companies spendings on advertisements is not readily available, though it would be a valuable piece of information to have, it is confidential and the companies have no obligation to disclose the same. It could also give the competitors an edge. Google-trends provides data on "interest over time" of the respective companies. This could be a good measure of share-of-voice for the company. This will be the independent predictor. Market share [46] [47] of company or similar data can be obtained as a monthly statistic for few years. This will be proportional data, assumed to follow a generalized Dirichlet distribution. This will be the prediction.

Regression problems based on compositional data can be categorized into 2 groups. In the first group the dependant variable is compositional [26], such problems have been solved by using Aitchison's geometric transformations. In the second group the response variables are compositional, with either same or different predictors. The latter type of problems is more complex and

this is what we have tackled in this chapter. So far, Dirichlet regression [48] has been the best approach in the industry to define compositional regression problems where the dependent variables are compositional.

In this chapter, we develop generalized Dirichlet (GD) regression, where the dependent variables follow a generalized Dirichlet distribution [49]. With double the number of parameters to estimate in comparison with the Dirichlet, GD is more versatile and gives space to model a flexible line of fit. We show how data fitting can be done by using a geometric transformation that reduces the generalized Dirichlet into a product of Beta distributions.

The rest of the chapter describes the main crux of our research, it will answer the pinching question, will the Google trends be able to predict the rise and fall of shares in any given field. To demonstrate the same, we have chosen the following three interesting quintessential markets of the modern world, related to technology and communication, mobile vendors in Canada, social network sites in India and what is the most used browser in the world! Let's explore this case and discuss how successful is Google-trends in predicting the trends of the share markets. Section 3.2 describes the machine learning algorithms employed for the research. Section 3.2.3 explains our contribution in devising the GD regression algorithm. Section 3.3 gives the background for experimental set-up. Section 3.4 shows our results and analysis. We conclude with Section 3.5. This work has been accepted as a conference paper [50].

3.2 Machine learning techniques

In this section we discuss the various ML algorithms used in our research. They are explained in the increasing order of computational complexity. Starting from ordinary least squares with a combination of transformations like *clr* and *ilr*. After which we describe Beta regression, Dirichlet regression and generalised Dirichlet regression.

3.2.1 OLS - Ordinary Least Squares Regression

Ordinary least squares (OLS) regression is a case of generalized linear modelling algorithm. It employs linear least squares method for estimating a single response variable. It could be multivariate of single independant variable x, to predict the response [51]. Principle of least squares minimizes the sum of squares of the differences between the actual response y, in the given data and prediction

of the linear function [52]. If we consider a linear system of variables, with *n* data points:

$$\sum_{j=1}^{n} X_{ij}\beta_j = y_i, (i = 1, 2, \ldots, m) \tag{23}$$

here β is the regression co-efficient

$$\mathbf{X}\beta = \mathbf{y} \tag{24}$$

where

$$\begin{gathered}\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \\ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}\end{gathered} \tag{25}$$

Such a system usually has no arithmetic solution, rather the aim is to find the best coefficients $\boldsymbol{\beta}$ which better fit the equations, to solve the quadratic minimization problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \tag{26}$$

$$\begin{gathered}\text{where the objective function } S \text{ is given by} \\ S(\beta) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2\end{gathered} \tag{27}$$

Finally, $\hat{\boldsymbol{\beta}}$ is the coefficient vector of the least-squares hyperplane, expressed as a product of Gramiam matrix of *X* and moment matrix of regressors:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{28}$$

3.2.2 Aitchison Transformations

In the share market case, both dependent and independent variables are compositional. Hence both are transformed to the Aitchison plane by applying the transformations explained below. Then, they are fed to ordinary least squares regression algorithm (ols). The resultant matrix is transformed back to Euclidean plane by applying an inverse transform. Then, the actual and predicted values are compared against the selection criteria explained in section 3.3.2. We discuss two Aitchison transformations [23], CLR (centred log ratio transform) and ILR (Isometric log ratio transform) in section 2.3 in Chapter 2. These are widely used in the case of compositional data.

3.2.3 Distributions-Based Regression

Beta regression

Assuming the response data is Beta distributed [53], The authors in [54] have proposed a regression model with mean and dispersion parameters of the distribution. In contrary to the transformed response of a linear regression, Beta regression's parameters are deduced using maximum likelihood estimation. The Beta density function is given as follows:

$$Y \sim \mathcal{B}(p,q), f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{29}$$

Maximum likelihood estimation is performed to deduce the values of p and q. The closed form solution to this equation is given in [54]. The partial derivatives of log of Beta distribution with respect to p and q are given by

$$\frac{\partial \log f(y;p,q)}{\partial p} = \psi(p+q) - \psi(p) + \log y \tag{30}$$

$$\frac{\partial \log f(y;p,q)}{\partial q} = \psi(p+q) - \psi(q) + \log(1-y) \tag{31}$$

where $\psi(.)$ is the digamma function defined as

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt \tag{32}$$

The expected score equals zero, it can be re-written as:

$$E[\log Y] = \psi(p) - \psi(p+q) \tag{33}$$

$$E[\log(1-Y)] = \psi(q) - \psi(p+q) \tag{34}$$

The distribution of response variable Y_i is $B(p_i, q_i)$ where p_i and q_i are, for each *i*, described by sets of explanatory variables $(x_1, \ldots, x_m)$ and $(v_1, \ldots, v_M)$ as

$$p_i = g(\beta_1 x_{1i} + \cdots + \beta_m x_{mi}) \tag{35}$$

$$q_i = h(\gamma_1 v_{1i} + \cdots + \gamma_M v_{Mi}) \tag{36}$$

Here g and h are link functions. The above equations can be substituted in the log likelihood equation of Beta distribution

$$\begin{aligned} \ell(\theta) = &\sum_{i=1}^{n} \log\Gamma(p_i + q_i) - \sum_{i=1}^{n} \log\Gamma(p_i) - \sum_{i=1}^{n} \log\Gamma(q_i) \\ &+ \sum_{i=1}^{n} p_i \log y_i + \sum_{i=1}^{n} q_i \log(1-y_i) \end{aligned} \tag{37}$$

whose first order derivatives are given by

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_r} &= \sum_{i=1}^{n} g_i' x_{ri} \left[\psi(p_i + q_i) - \psi(p_i) + \log y_i\right] \\ \frac{\partial \ell}{\partial \gamma_R} &= \sum_{i=1}^{n} h_i' v_{Ri} \left[\psi(p_i + q_i) - \psi(q_i) + \log(1-y_i)\right] \end{aligned} \tag{38}$$

Maximum likelihood estimation of β and γ are obtained by solving the above equations, equating to zero. Thus, the regression parameters are obtained. They can be multivariate or univariate, depending on the application. It has to be noted that, in the case of compositional data, where the predicted values are more than one, Beta regression needs to be extended to accommodate the prediction of multiple dependant variables. This is explored in the further two sections, Dirichlet regression and generalized Dirichlet regression.

Dirichlet regression

Maier has proposed Dirichlet regression [48] [55], which assumes dependent variables are compositional and follow a Dirichlet distribution. He has deduced a framework similar to general linear models for regression of Dirichlet distributed data. Dirichlet distribution is a generalized form of Beta distribution [10], defined in equation 39. Also known as common parametrization.

$$\mathcal{D}(y|\alpha) = \frac{1}{\mathrm{B}(\alpha)} \prod_{c=1}^{C} y_c^{(a_c-1)} \tag{39}$$

$$\text{where } B(\alpha) = \prod_{c-1}^{C} \Gamma\left(\alpha_c\right) / \Gamma\left(\sum_{c-1}^{C} \alpha_c\right) \tag{40}$$

$$\text{and } \Gamma(x) = \int_0^{\infty} t^{x-1} \exp(-t) \mathrm{d}t \tag{41}$$

Alternately, Dirichlet distribution can also be represented as a function of mean μ and variance ϕ as in equation 42, called alternate parametrization.

$$f(y|\mu, \phi) = \frac{1}{\mathrm{B}(\mu\phi)} \prod_{c=1}^{C} y_c^{(\mu_c\phi-1)} \tag{42}$$

The full log-likelihood of the commonly parametrized model is defined below

$$\ell_c(y|\alpha) = \log\Gamma\left(\sum_{c=1}^{C} \alpha_c\right) - \sum_{c=1}^{C} \log\Gamma\left(\alpha_c\right) + \sum_{c=1}^{C} \left(\alpha_c - 1\right) \log\left(y_c\right) \tag{43}$$

The crucial part of converting a Dirichlet distribution to a Dirichlet regression problem, lies in the link between the Dirichlet parameters (α) and the regression parameters (β). The link function *g(.)* is selected as a *log(.)* function, defined as

$$g(\alpha_c) = \eta_c = \boldsymbol{X}^{[c]} \boldsymbol{\beta}^{[c]} \tag{44}$$

The first order derivative of the log-likelihood:

$$\frac{\partial \ell_c}{\partial \beta_m^{[d]}} = x_m^{[c]} \alpha^{[d]} \left[\log\left(y^{[d]}\right) - \psi\left(\alpha^{[d]}\right) + \psi\left(\sum_{c=1}^{C} \alpha^{[c]}\right) \right] \tag{45}$$

The second order derivatives of the log-likelihood with respect to βs on the same and different

variables are given below. The Hessian matrix of the same can be obtained from [56].

$$\frac{\partial^2 \ell_c}{\partial \beta_m^{[d]} \partial \beta_n^{[d]}} = x_m^{[d]} x_n^{[d]} \alpha^{[d]} \left\{ \log\left(y_d\right) + \psi\left(\sum_{c=1}^{C} \alpha_c\right) - K \right\} \tag{46}$$

$$K = \psi\left(\alpha_d\right) - \alpha_d \left[\psi_1\left(\sum_{c=1}^{C} \alpha_c\right) - \psi_1\left(\alpha_d\right) \right] \tag{47}$$

$$\frac{\partial^2 \ell_c}{\partial \beta_m^{[d]} \partial \beta_n^{[e]}} = x_m^{[d]} x_n^{[e]} \alpha_d \alpha_e \psi_1\left(\sum_{c=1}^{C} \alpha_c\right) \tag{48}$$

Caution is to be taken to resist the urge to calculate α of the Dirichlet distributed dependent variables. As this is not the desired output, we are more interested in calculating the regression parameters β, which will be found only after relating it to α in the link function. The maximum log-likelihood estimation (MLE) of Dirichlet regression is different from that of Dirichlet distribution MLE.

Generalized Dirichlet (GD) regression

We now propose generalized Dirichlet regression to solve the compositional regression problems of interest. This distribution has double the number of parameters to estimate compared to Dirichlet distribution. This gives it more degrees of freedom to fit the data in a better way. It's probability density function has the following form:

$$c \prod_{i=1}^{n} \left[x_i^{a_i - 1} \left(1 - \sum_{k=1}^{i} x_k\right)^{b_i - 1} \right] \tag{49}$$

To evaluate the normalizing constant *c*, we integrate sequentially over $x_n, x_{n-1}, \ldots, x_2, x_1$, where *n* is the number of components of *x*.

$$c = \prod_{i=1}^{n} \frac{\Gamma\left(1 + \sum_{k=i}^{n} \left(a_k + b_k - 1\right)\right)}{\Gamma\left(a_i\right) \Gamma\left(b_i + \sum_{k=i+1}^{n} \left(a_k + b_k - 1\right)\right)} \tag{50}$$

In an *n* dimensional GD distributed data, there are *2n* variables to estimate. There would be *2n* log-likelihood equations to solve simultaneously. [57] demonstrates that GD can be transformed to *n* Beta distributions. Suppose $\left(X_1, \ldots, X_n\right) \sim GD\left(a_1, \ldots, a_n; b_1, \ldots, b_n\right)$ where $Zi = Zi...Zn$

are *n* mutually independent Beta distributed variables

$$Z_i \sim B\left(a_i, b_i + \left(\sum_{k=i+1}^{n} \left(b_k + a_k - 1\right)\right), i = 1, \ldots, n \right. \tag{51}$$

$$\left(X_1, \ldots, X_n\right) \triangleq \left(Z_1, Z_2\left(1 - Z_1\right), \ldots, Z_n \prod_{i=1}^{n-1}\left(1 - Z_i\right)\right) \tag{52}$$

It follows that the problem can be reduced to *n*-likelihood equations in pairs.

$$Z_i = X_i / \left(1 - \sum_{j=1}^{i-1} X_j\right) \sim B\left(a_i, c_i\right) \tag{53}$$

where $i = 1, ..., n$ for a random sample $\left(X_{1j}, \ldots, X_{nj}\right), j = 1, \ldots, N,$ from $X_1, \ldots, X_n$ the corresponding log-likelihood function can be expressed as follows:

$$\prod_{l=1}^{N} \prod_{i=1}^{n} \left[\frac{\Gamma\left(a_i + c_i\right)}{\Gamma\left(a_i\right)\Gamma\left(c_i\right)} \left(\frac{x_{il}}{1 - \sum_{j=1}^{i-1} x_{jl}}\right)^{a_i - 1} \left(\frac{\sum_{j=1}^{i} x_{jl}}{1 - \sum_{j=1}^{i-1} x_{jl}}\right)^{c_i - 1}\right] \tag{54}$$

The first order derivatives of MLE are *n* pairs, where $i = 1, ..., n$ is below:

$$\frac{\partial \log L}{\partial a_i} = \psi\left(a_i\right) - \psi\left(a_i + c_i\right) - \frac{1}{N} \sum_{l=1}^{N} \log x_{il} = 0 \tag{55}$$

$$\frac{\partial \log L}{\partial c_i} = \psi\left(c_i\right) - \psi\left(a_i + c_i\right) - \frac{1}{N} \sum_{l=1}^{N} \log\left(1 - x_{il}\right) = 0 \tag{56}$$

Since a Dirichlet MLE has been transformed to *n* Beta MLE, we will follow the steps in section 3.2.3 to convert the Beta distribution to Beta regression estimates using link function. Various link functions can be used to relate the dependent variables to independent variable, for example, log-link, logit-link, probit, log-log link.

There is no closed form solutions for the above equations. Newton-Raphson iteration is employed to arrive at the solution in maximum likelihood estimation. The initial values are obtained from method of moments estimates [58]. The estimated values of regression are normalised to equate to 1 as the expectation is compositional data.

3.3 Experimental Set-up

3.3.1 k-fold cross validation

Our aim is to create a replica of the underlying model generating the data. The approach we follow is to work backwards from the collected data. In the process, there is a danger of over-fitting/under-fitting the data. This means that, the model is created specifically around the given data. Any newly generated data, even though coming from the same source, will not be described with this model. To overcome this issue, we have used the method of k-fold cross validation [59]. It is a systematic hold out method, where the data is splitted into k parts. over a loop of k times, each part is held out for testing and the model is trained over the remaining data (total-kth part). This way k different models are created, and the features are averaged over the k models. This gives equal opportunity for the data being represented fully compared to randomized hold-out cross validation.

The k-fold algorithm is used to compute the average of evaluation measures, and it is called 1000 times to average out the measures over the different partitions. Thus, each dataset is modelled 1000*10 times, that is 10,000 times for a 10-fold cross validation. This is the most computationally expensive component of the regression problem we are solving.

Algorithm 1 k-fold cross-validation pseudo-code

Input Dataset, number of folds k,
Split the data into k equally sized (rounded to integer) folds

```
1: procedure K-FOLD-REGRESSION
2:     for <s = 1:k> do
3:         Train a model on sth fold's training set
4:         Test the model on sth fold's test set
5:         Compute corresponding evaluation measures
6:     end for
7:     Compute average of evaluation measures of all k sets
8: end procedure
```

3.3.2 Evaluation Measures

The goodness of fit of the regression models needs to be measurable. This helps us decide which model is able to describe the data best. Since we are dealing with compositional data, the regular measures of regression need to be modified accordingly, There are various measures explained in [60]. The efficacy of the learned models is compared with these three parameters, sum of square of

residuals, R-squared measure based on total variability and KL divergence.

R-squared measure based on total variability (R2T)

Th term total variability based R square measure was coined by Aitchison in the log ratio analysis [23]. It is defined as the ratio of total variance in the predicted values to the total variance in the actual values.

$$R_T^2 = \text{totvar}(\widehat{\mathbf{x}}) / \text{tot}\, var(\mathbf{x}) \tag{57}$$

$$\mathbf{T}(\mathbf{x}) = [\tau_{ij}] = [\text{var}\left\{\log\left(x_i/x_j\right)\right\}] \tag{58}$$

$$\text{totvar}(\mathbf{x}) = \frac{1}{2d}\sum \mathbf{T}(\mathbf{x}) \tag{59}$$

Residual Sum of Squares (RSS)

Residual sum of squares (RSS) [61] is defined as the square of the difference between the predicted and actual values for each point in test set. In the compositional case, the sum of each component's RSS is summed up.

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{60}$$

Kullback-Leibler divergence (KL)

KL divergence is the sum of ratio of the logarithm of actual values of fitted values, weighed by the actual, over each data point [62]. Minimum KL divergence is desired as it deduces maximum likelihood [63].

$$KL(\mathbf{S}, \widehat{\mathbf{S}}) = \sum_{t=1}^{T}\sum_{j=1}^{D} \log\left(\frac{S_{jt}}{\widehat{S}_{jt}}\right) S_{jt} \tag{61}$$

KL divergence adapted to compositional data is defined as follows [64]:

$$KLC(\mathbf{S}, \widehat{\mathbf{S}}) = \frac{D}{2}\left(KL\left(\mathbf{0_D}, \mathbf{S} \ominus \widehat{\mathbf{S}}\right) + KL\left(\mathbf{0_D}, \widehat{\mathbf{S}} \ominus \mathbf{S}\right)\right) \tag{62}$$

$$= \frac{D}{2}\sum_{t=1}^{T} \log\left(\overline{\left(S_t/\widehat{S}_t\right)} \cdot \overline{\left(\widehat{S}_t/S_t\right)}\right) \tag{63}$$

3.4 Datasets and Results

In order to assess the usefulness of the regression models, we have investigated 2 real-life data sets followed by 3 different applications based on data collected from real-life sources. The market shares data are obtained from global-stats[1] website, the relation we are trying to observe is the company's market-share to their trends in google-searches[2], which is a good measure of the company's investment in advertising.

3.4.1 Real Data

Arctic Lake soil compositions dataset

We discuss Arctic lake data set, which shows how the composition of ground soil comprising of silt, sand and clay is altered as the depth of lake increases. This is a famous dataset used by Aitchison to investigate many transformations. It has been quoted in studies, like zero value substitution [65] [66] and robustness checks [66]. There are 39 data points with three components. The distribution of the data points is shown in ternary diagram, Figure 3.1. Unlike most regression problems, with a single response variable and multiple independent variables, the Arctic lake data is different. It has a single independent variable (x) and three different response variables, whose sum adds up to unity. It is the compositions that we are predicting. We could use a variety of arithmetic transformations of x, like log(x), x_2, x + x_2. Table 3.1 shows the regression measures of the different regression algorithms explained in the above sections. GD regression has least residual sum of squares, which represents a good fit. Dirichlet regression is the best in this case, given R2T and KL divergence measures.

Glass composition dataset

Another dataset containing more number of compositions is the forensic glass dataset [65]. It has 8 components and 214 observations. They are all dependant on a single parameter, the refractive index (RI) of glass. The aim is to map how the refractive index of glass can alter the composition of glass, the minerals in it, like Aluminium, magnesium to name a few. The regression should ideally appear the other way round, the RI is determined by the composition. But, we are doing the

[1] http://gs.statcounter.com
[2] https://trends.google.com/trends/

reverse process to see if we can find the required composition to be able to manufacture/recreate the intended RI. As per the results in Table 3.2, least SSR is by GD regression algorithm. Better KL divergence is shown by the classical OLS methods, this could be due to the unequal distribution of metals in the glass. Some metals like Silicon and Sodium have high compositions compared to the rest.

3.4.2 Application - Market Shares for Information Technology Companies

For future researchers to be able to reproduce our work, we have chosen the public platform of google-trends [67] [68]. It gives us a very good idea on how the data search has spiked over the given time range, which is of prime importance. Here have been a couple of experiments done to see if any lag in trends and share is observed. Trends seem to be more real time and the data seemed to be more relevant to the current market share. A lag of two months is observed between the trends and market shares. The companies investment on advertisements seem to have been fruitful a couple of months later in getting the google clicks and thus for it to show affect on the market share. We would further like to explain how the shares have changed over the time, any interesting patterns are explained. It is to be noted that google-trends [69] only supports comparison of 5 key-words at a time. To support more searches, we will compare the term with a standard term such as "photo" to get a relative measure of frequency. This process is done with all the variables, then put together and normalized.

Browser market shares - Worldwide

It is interesting to note, in 2009, Internet-Explorer (64.97%) and Firefox (26.85%) were leading the market, and today they are mere 3-5% share holders in the world-wide browser markets. This owes to the introduction of new browser by Google, Chrome and Apple's Safari. The major market shares in internet browsers is held by Chrome (51.5%) followed by Safari (15.13%) in 2018. We have collected monthly from 2009 January to 2018 September, with $n = 118$ data points, with $D = 6$ components, including UC browser and Opera. The ternary diagrams of 3 sets of browsers are given in Figures 3.2, 3.3, 3.4 and 3.5. This shows how the compositions have changed over the course of 10 years. They clearly follow a linear pattern. Table 3.3 displays the results of the experiment. It is observed that GD regression shows close to unity R2T, and least KL divergence (0.23). Dirichlet regression has slightly better RSS (0.03) than GD regression (0.09). The arithmetic

Table 3.1: Generalized Dirichlet Regression measures of Arctic lake sediments data

Methods\Measures	SSR	R2T	KL
CLR + OLS	3.3287	11.0777	25.4432
ILR + OLS	3.3183	11.8528	25.3671
Dirichreg	0.8193	6.2170	15.9382
Generalized Dirichlet	0.5415	8.9557	17.9848

Table 3.2: Generalized Dirichlet Regression measures of forensic glass data

Methods\Measures	SSR	R2T	KL
CLR + OLS	9.4718	0.0126	216.8152
ILR + OLS	9.4577	0.0129	217.3527
Dirichreg	0.0186	0.0023	333.3006
Generalized Dirichlet	0.0145	0.0015	336.3690

transforms combined with OLS with less computational complexity, take lesser time to execute, with acceptable results.

Mobile Seller market shares in Canada

The Canadian mobile market was ruled by Apple (88.97%) in 2010. It is now sharing space with Samsung (25.14%) in 2018. We have included LG, Huawei, Google and Motorola in the study. With $D = 6$ and $n = 95$, we have the independent variables obtained from Google-trends, individually for each company, the trends are in accordance with the share market patterns. Table 3.4 describes the results. Looking at the measures, we can say that google trends has been a good measure of predicting the shares, with the regression fits of RSS close to zero. The generalized Dirichlet regression performed well compared to other methods.

Social Networks market shares in India

Facebook is the most followed social networking site in India. It has been growing popularity from 52.3% in 2010 to 86.56% in October 2018. Many companies have mushroomed in this space but Youtube has sustained it's second place with 10% and it saw it's peak with 25% in 2012. Twitter had a good 7% share in 2013, but now it has a mere 1% share. Results have been recorded in Table 3.5. Dirichlet regression seems to fit the data better than GD regression, with slight variation in the measures.

Table 3.3: Generalized Dirichlet Regression measures of world-wide browser shares

Methods\Measures	SSR	R2T	KL
CLR + OLS	4.1566	0.9075	0.7468
ILR + OLS	6.2981	0.0386	11.4344
Dirichreg	0.0279	1.0533	0.2290
Generalized Dirichlet	0.0840	0.8230	0.8385

Table 3.4: Generalized Dirichlet Regression measures of mobile vendor shares in Canada

Methods\Measures	SSR	R2T	KL
CLR + OLS	4.9768	5.6769	9.6099
ILR + OLS	4.9934	0.3337	9.1590
Dirichreg	0.0004	0.9443	0.0044
Generalized Dirichlet	0.0100	0.2747	0.1194

Table 3.5: Generalized Dirichlet Regression measures of Social Media Shares in India

Methods\Measures	SSR	R2T	KL
CLR + OLS	5.0926	0.0419	11.3324
ILR + OLS	5.1004	0.0395	11.4662
Dirichreg	0.0720	0.0741	11.5216
Generalized Dirichlet	0.6679	0.1262	21.2748

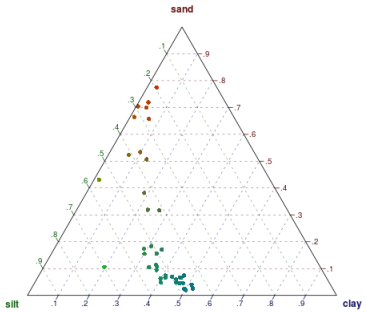


Figure 3.1: Arctic Lake

3.5 Conclusion

We have introduced an implementation of generalized Dirichlet regression that extends Beta regression for compositional, multiple response variables. An application in share-market analysis demonstrates the modelling capabilities of this solution. Various compositional regression models have been discussed, and their results compared. The question, ”Is google-trends a good predictor of the share market dynamics?” is answered with three real-world examples. Google trends seem to

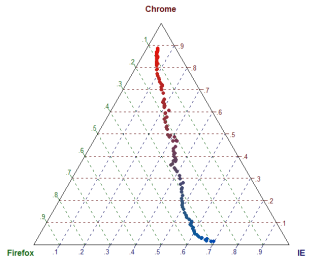


Figure 3.2: Browser shares: Chrome, IE, Firefox

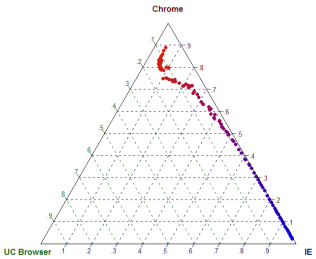


Figure 3.3: Browser shares: Chrome, IE, UC

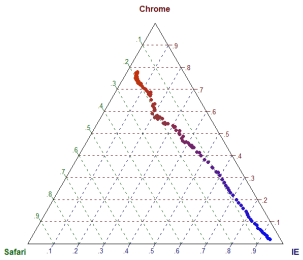


Figure 3.4: Browser shares: Chrome, Safari, IE

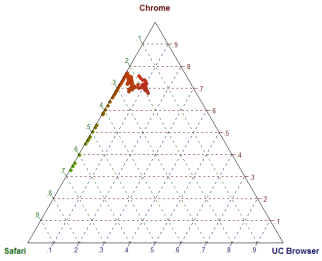


Figure 3.5: Browser shares: Chrome, UC, Safari

capture the share-market trends well. The distribution-based regression algorithms fared better than transformations-based regression. Though the trade-off is the use of more computationally complex calculations.

Additionally, we suggest future work should explore new choices of starting values for the maximum likelihood estimation of generalized Dirichlet regression and their corresponding sensitivity studies for these choices. A more robust system with less sensitivity to the estimation of regression coefficients could be developed, with the use of mixture models in Dirichlet regression.

Chapter 4

Beta-Liouville Regression and Applications

4.1 Introduction

Compositional data [23] [70] have been studied in metabolomics [1], hematology [48], electoral votes [53]. Since Beta regression has been introduced by [54] [71], it has been vastly used for the study of regression in compositional data. However, it is limited to one dimension. A generalization of Beta regression was proposed as Dirichlet regression [48] [55]. A more generic form of distribution based regression is proposed in this chapter, the Beta-Liouville regression for continuous and compositional/proportional data. Beta-Liouville is a part of the Liouville family of distributions. It has been studied intensively in classification and clustering problems in computer vision, pattern recognition and image processing [72] [73] [74] [56] [75].

We had several goals in carrying out this research. The first was to investigate the Liouville family of distributions in regression analysis of compositional data and apply it to the fields of market-share analysis and occupancy estimation of smart buildings. The Beta-Liouville (BL) distribution has a more general covariance structure compared to Dirichlet, which makes it more useful in real-life applications.

We want to predict the market-share [46] [43] of a company that has started offering its shares of stock to the public. Market share is the percentage of total shares of a given company for that particular product in the market, whereas shares of stock is part of the company's ownership which

is valued as a stock certificate. Following the general pipeline of machine learning, we first gather a training set of data from all the companies manufacturing similar products in the given region with known market shares. The region could mean a specific country, province or even the world wide market. Next, we need to design features that are relevant to the task. The company's investment in marketing is one of potential features. We expect that higher the marketing expenditure, more market-share the company owns. However, the money a company spends on its advertisements is confidential. So a close approximation to it would be the internet presence of the company. In the post-internet era today, the amount of discussion about a product on social networks like Facebook, Twitter and even Google searches have shown to reliably predict the popularity of the brand. Here the input feature is the Google-trends [67] data on how trending the company is on a certain month, the output being its market-share during the corresponding time period. Now, in order to connect the market-share to the marketing expenditure, we train a linear model or regression line using our training data. Once the model is trained, the market share of a company can be predicted based on its Google trend. Finally, comparing the predicted share to the actual share for a testing set of data, we can test the performance of our regression model.

Another application explored in this chapter is inspired from [76], estimating occupancy in smart buildings. Gaining information on the number of occupants in a given room helps us determine the ambient living conditions. Accurate estimation allows in turn to set the automatic thermostat or switches off unnecessary lights in a smart building, saving money and resources. It is costly to depend on cameras or counting gates to gain information of occupancy for each room in a large building/home/office. However, with the use of existing simple everyday apparatus, we can gain few insights and use it to predict the number of occupants. This is a cost-effective method. There have been several studies regarding smart buildings in the past.

This chapter describes the Liouville family of distributions, Beta distribution and Beta-Liouville model in section 4.2. The parameter estimation and link functions are explained in section 4.3. Section 4.4 delineates a set of challenging applications and presents our results. We conclude with Section 4.5. This work has been accepted as a conference paper [77].

4.2 The Model

In this section we shall discuss the Liouville family of distributions, Beta-Liouville distribution and its relation to Dirichlet distribution.

4.2.1 Liouville Family of Distributions

For a vector $Y = (Y_1, ..., Y_D)$ with D variate Liouville distribution with positive parameters $(\alpha_1, ..., .\alpha_D)$ and density generator $g(.)$, the probability density function is given by

$$p(\boldsymbol{Y}|\alpha_1, \ldots, \alpha_D) = g(u) \prod_{d=1}^{D} \frac{Y_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \tag{64}$$

where $u = \sum_{d=1}^{D} Y_d < 1, Y_d > 0, d = 1, \ldots, D$ The general moment function of a Liouville distribution [78] is

$$E\left(Y_1^{r_1} Y_2^{r_2} \ldots Y_D^{rd}\right) = E\left(U^r\right) \frac{\prod_{d=1}^{D} \Gamma(\alpha_d + r_d) \Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{\prod_{d=1}^{D} \Gamma(\alpha_d) \Gamma\left(\sum_{d=1}^{D} \alpha_d + r\right)} \tag{65}$$

The mean, variance and covariance are as follows.

$$E(Y_d) = E(U) \frac{\alpha_d}{\sum_{d=1}^{D} \alpha_d} \tag{66}$$

$$\begin{aligned} \operatorname{Var}(Y_d) = E\left(U^2\right) & \frac{\alpha_d(\alpha_d + 1)}{\sum_{d=1}^{D} \alpha_d \left(\sum_{d=1}^{D} \alpha_d + 1\right)} \\ & - E(Y_d)^2 \frac{\alpha_d^2}{\left(\sum_{d=1}^{D} \alpha_d\right)^2} \end{aligned} \tag{67}$$

$$\operatorname{Cov}(Y_l, Y_k) = \frac{\alpha_l \alpha_k}{\sum_{d=1}^{D} \alpha_d} \left(\frac{E\left(U^2\right)}{\sum_{d=1}^{D} \alpha_d + 1} - \frac{E(U)^2}{\sum_{d=1}^{D} \alpha_d} \right) \tag{68}$$

The r^{th} moment of a random variable $U \in [0, 1]$ is given by $E(U^r)$, where $r = r_1 + \cdots + r_d$. Its probability density function $f(.)$ is related to the density generator $g(.)$ as follows

$$g(u) = \frac{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}{u \sum_{d=1}^{D} \alpha_d - 1} f(u) \tag{69}$$

Substituting the value of $g(u)$ in Eq. 64, the Liouville distribution of the second kind can be written as follows

$$p(\boldsymbol{Y}|\alpha_1,\ldots,\alpha_D) = \frac{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)}{u\sum_{d=1}^{D}\alpha_d - 1} f(u)\prod_{d=1}^{D}\frac{Y_d^{\alpha_d-1}}{\Gamma\left(\alpha_d\right)} \tag{70}$$

4.2.2 Beta-Liouville Distributions

Beta distribution is a good choice for the distribution of u. The density function of Beta distribution with positive parameters α and β is given as follows

$$f(u|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}u^{\alpha-1}(1-u)^{\beta-1} \tag{71}$$

with mean and variance

$$E(u) = \frac{\alpha}{\alpha+\beta} \quad E\left(u^2\right) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \tag{72}$$

$$\text{Var}(u) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{73}$$

To obtain Beta-Liouville [79] distribution, we substitute Eq. 71 in Eq. 70

$$\begin{aligned} p(\boldsymbol{Y}|\alpha_1,\ldots,\alpha_D,\alpha,\beta) = \frac{\Gamma\left(\sum_{d=1}^{D}\alpha_d\right)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \\ \times\prod_{d=1}^{D}\frac{Y_d^{\alpha_d-1}}{\Gamma(\alpha_d)}\left(\sum_{d=1}^{D}Y_d\right)^{\alpha-\sum_{d=1}^{D}\alpha_d}\left(1-\sum_{d=1}^{D}Y_d\right)^{\beta-1} \end{aligned} \tag{74}$$

The mean, variance and covariance of Beta-Liouville distribution is obtained by substituting Eq 72 in Eq. 66, 67 and 68.

$$E\left(Y_d\right) = \frac{\alpha}{\alpha+\beta}\frac{\alpha_d}{\sum_{d=1}^{D}\alpha_d} \tag{75}$$

$$\begin{aligned} \text{Var}\left(Y_d\right) = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}\frac{\alpha_d\left(\alpha_d+1\right)}{\sum_{d=1}^{D}\alpha_d\left(\sum_{d=1}^{D}\alpha_v+1\right)} \\ - \frac{\alpha^2}{(\alpha+\beta)^2}\frac{\alpha_d^4}{\left(\sum_{d=1}^{D}\alpha_d\right)^4} \end{aligned} \tag{76}$$

$$\begin{aligned}\operatorname{Cov}\left(Y_l, Y_k\right) =& \frac{\alpha_l \alpha_k}{\sum_{d=1}^{D} \alpha_d}\left(-\frac{\alpha^2}{(\alpha+\beta)^2 \sum_{d=1}^{D} \alpha_d}\right. \\ &\left.+\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)\left(\sum_{d=1}^{D} \alpha_d+1\right)}\right)\end{aligned} \tag{77}$$

It is interesting to note that Beta-Liouville distribution can be reduced to a Dirichlet distribution of $D+1$ components when we substitute $\alpha=\sum_{d=1}^{D} \alpha_d$ and $\beta=\alpha_D$:

$$f(u)=\frac{\Gamma\left(\alpha_{D+1}+\sum_{d=1}^{D} \alpha_d\right)}{\Gamma\left(\alpha_{D+1}\right) \Gamma\left(\sum_{d=1}^{D} \alpha_d\right)} u^{\sum_{d=1}^{D} \alpha_d-1}(1-u)^{\alpha_{D+1}} \tag{78}$$

4.3 Model Learning

4.3.1 Link Function

The inverse of any cumulative distribution function corresponding to a continuous distribution is called the link function [53] [49] relating the regression parameters to covariates. The parameters of Beta-Liouville distribution can be related to the covariates to form a regression line as follows

$$\begin{aligned}\alpha &= g_1\left(\boldsymbol{\beta}_1 x_1+\cdots+\boldsymbol{\beta}_m x_m\right) \\ \beta &= h\left(\boldsymbol{\gamma}_1 v_1+\cdots+\boldsymbol{\gamma}_M v_M\right) \\ \alpha_d &= g_2\left(\boldsymbol{\beta'}_{1d} x_{1d}+\cdots+\boldsymbol{\beta'}_{md} x_{md}\right)\end{aligned} \tag{79}$$

we have chosen the link functions to be $g_1(.)=h(.)=logit(.)$ and $g_2(.)=log(.)$ If we consider the final regression equation $Y=m_0+m_1 x_1+...+m_n x_n$, the relationship between the regression parameters of $\boldsymbol{\beta}$ and $\boldsymbol{\beta'}$ to m is given as $m=\boldsymbol{\beta}\boldsymbol{\beta'}D$, where D is the dimensionality of Y.

4.3.2 Parameter Estimation

Maximum likelihood estimation (MLE) [58] [80] is chosen as there is no closed-form solution to Beta-Liouville regression. MLE is a process that tries to find the most likely model to have produced the observed outcome. The mechanics of the algorithm involve a Newton-Raphson optimization that iterates between a scoring step, based on the current parameters, and an update to the parameters

to improve the fit. The log-likelihood function $\ell(\theta)$ is described with a joint parameter vector $\theta = \left(\boldsymbol{\beta}^T, \boldsymbol{\beta'}^T, \boldsymbol{\gamma}^T\right)^T$. The first derivatives of log-likelihood equation with respect to regression parameters are given as follows

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} = g' x \Bigg[& \Psi\left(\alpha+\beta\right) - \Psi\left(\alpha\right) \\ & + \sum_{i=1}^{N} \log \sum_{d=1}^{D} Y_{id} \Bigg]\end{aligned} \tag{80}$$

$$\begin{aligned}\frac{\partial l}{\partial \gamma} = h' v \Bigg[& \Psi\left(\alpha+\beta\right) - \Psi\left(\beta\right) \\ & + \sum_{i=1}^{N} \log \sum_{d=1}^{D} \left(1 - Y_{id}\right) \Bigg]\end{aligned} \tag{81}$$

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta'_d}} = x_d \alpha_d \Bigg[& \Psi\left(\sum_{d=1}^{D} \alpha_d\right) - \Psi\left(\alpha_d\right) \\ & + \sum_{i=1}^{N} \left(\log Y_{id} - \log \sum_{d=1}^{D} Y_{id}\right) \Bigg]\end{aligned} \tag{82}$$

where $\Psi(\alpha) = \frac{\partial \log \Gamma(\alpha)}{\partial \alpha}$ is the digamma function.

Newton-Raphson method is employed to estimate the new values of model parameters. The Hessian matrix has a block diagonal matrix format, shown as

$$H\left(\theta\right) = blockdiag\left(H_a, H_b\right) = \begin{bmatrix} H_a & 0 \\ 0 & H_b \end{bmatrix} \tag{83}$$

where

$$H_a = H\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \begin{bmatrix} \frac{\partial l}{\partial^2 \boldsymbol{\beta}} & \frac{\partial l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}} \\ \frac{\partial l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}} & \frac{\partial l)}{\partial^2 \boldsymbol{\gamma}} \end{bmatrix} \tag{84}$$

$$H_b = H\left(\boldsymbol{\beta'}_1, \ldots, \boldsymbol{\beta'}_D\right) = \frac{\partial l}{\partial \boldsymbol{\beta'}_{d_i} \partial \boldsymbol{\beta'}_{d_k}} \tag{85}$$

where $i, k \in \{1, \ldots, D\}$ The inverse of Hessian matrix can be computed as

$$\begin{aligned} H^{(-1)}\left(\theta_j\right) &= \left(block\,\mathrm{diag}\left(H_a, H_b\right)\right)^{(-1)} \\ &= blockdiag\left(\left(H_a\right)^{(-1)}, \left(H_b\right)^{(-1)}\right) \end{aligned} \tag{86}$$

The inverse of Hessian matrix is used in the Newton-Raphson equation to estimate new parameters.

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \mathbf{H}^{-1} \frac{\partial l}{\partial \boldsymbol{\theta}} \tag{87}$$

4.3.3 Initialization of parameters

Numericl maximization of the log-likelihood is required as there is no closed form solution in this case. Maximum likelihood estimation is used to estimate the parameters. This requires the specification of initial values to be used in the iterative scheme. The authors in [54] suggest the use of ordinary least squares estimate of the β parameter vector obtained from a linear regression of the transformed response on X.

As discussed in Eq. 78 Beta-Liouville distribution of second kind can be reduced to Dirichlet distribution [81] [82] with a specific set of parameters. Thus, the same method used in the case of Dirichlet can be used for the initialization of the Beta-Liouville method [75].

4.4 Datasets and Results

The primary purpose of this section is to apply the proposed Beta-Liouville regression and outline its effectiveness when compared to previously proposed techniques, namely ordinary least squares (OLS) regression and Dirichlet regression. As we have considered compositional data, the data is transformed using centered log ratio (clr) transform and isometric log ratio (ilr) transform before feeding to OLS algorithm. Details of clr, ilr, OLS and Dirichlet regression are well documented in [55] [48] [65] [83], for instance. We empirically test our approach on several real world applications in order to show its general capability and performance in various situations. Interesting applications included in this chapter are share-market analysis based on google-trends and smart-building occupancy estimation based on heterogenous sensors [76].

4.4.1 Real Data

We have investigated 2 famous real-life data sets traditionally used in compositional regression analysis, Arctic lake soil and forensic glass compositions.

Arctic Lake

We predict how the proportions of sand, silt and clay vary in Arctic lake ground soil with increase in depth of lake. Table 4.1 displays the results of different regression algorithms. As the depth of the lake increases, the percentage of sand decreases, clay and silt increases, in accordance to our own experience in the real-world. With $n = 39$ and $D = 3$ the small size of dataset is compensated with k-fold cross validation to justify the results. Beta-Liouville and Dirichlet regression perform well in comparison to ordinary least squares with data transformation.

Glass Compositions

The forensic glass dataset shows how the composition of glass varies with the refractive index of glass. The glass comprises of mainly Silicon and Sodium, with traces of Magnesium, Aluminium, Potassium, Calcium, Barium and Iron. The dataset has 8 components and 214 observations. Results of regression are recorded in Table 4.2. In all the methods, KL divergence is very high, this could be due to the extremely uneven distribution of parameters. Nonetheless the RSS and R2T measures of Dirichlet and Beta-Liouville regression are satisfactory, indicating a good fit.

4.4.2 Market Shares

In this section, we explore how the market-share dynamics of information technology companies change over a period of 10 years. We make a regression between Google-trends for the company name at the given point of time and their market-shares. We have chosen Browsers, mobile phones and social networks as the field of study of market-shares individually. The data is readily available at Global-stats[1] website and Google-trends[2] for those interested in reproducing this work.

Worldwide Browser market shares

In the year 2018, the worldwide browser market is shared by Google Chrome (61.5%), Safari (15%), UC browser (4.4%), Firefox (5%), Opera (3%) and Internet Explorer (2.8%). This was not the case in 2006 when Internet Explore and Firefox were leading with 65% and 27% shares respectively. It is interesting to note how the markets have changed and made space for our new browsing patterns. Google-trends have been proved to be a very good estimate for the share market.

[1] http://gs.statcounter.com
[2] https://trends.google.com/trends

A good fit of regression line is proof to this. The data set used here includes 6 components and 118 data-points. Table 4.3 records the results of regression. With very low residuals and KL divergence and close to unity R2T measure, Beta-Liouville regression has proves to be a good fit.

Canadian Mobile Phone market shares

Canadians have been ardent fans of the mobile phone company Apple (88.97%) since 2010. Recently a good portion of the users have switched to Samsung, raising its shares to 25% in 2018. Other competitors include LG, Huawei, Google and Motorola. The dataset includes 6 components and 95 data-points. Table 4.4 displays the results for this data.

India's Social Network market shares

In India, Facebook has been the the most captivating social network, consistently improving its market-share from 50% to 86% over a period of 8 years. There have been quiet a number of companies mushrooming in this space, eg. Orkut, Digg, StumbleUpoon, Mixx etc. However only Facebook, Youtube and Twitter have stood the test of time. The dataset consists of 6 components and 106 instances, collected monthly from 2010-18. Results are shown in Table 4.5.

4.4.3 Smart buildings

Building energy consumption is majorly influenced by occupant behaviour. Control systems and modeling methods to assist occupants have been discussed in [84]. Occupant behaviour models are built using electricity metered office appliance data. Occupant's computer logs the arrival/departure of occupant. This could be intrusive and bothersome. Whereas in our study we avoid it with available non-intrusive sensors [76]. There have been various studies in multi-sensor and multi-feature models to esimate occupancy levels. Studies in [85] [86] have used passive infra-red (PIR), motion and acoustic sensors. Magnetic reed switches and motion sensors used to estimate occupancy, double utilize the HVAC systems in smart buildings.

The data are gathered from an office set up in Grenoble Institute of Technology, accommodating a professor and 3 students. The office has frequent visitors for meetings and presentations all week long. The ambience sensing network set-up gives information on the power consumption, motions, acoustic pressure decibel from microphone, door and window position. This is used to estimate the number of occupants in the office. The actual measure of the number of people in the room is taken

Table 4.1: Beta-Liouville Regression measures of Arctic lake sediments data

Methods\Measures	RSS	R2T	KL
CLR + OLS	3.3287	11.0777	25.4432
ILR + OLS	3.3183	11.8528	25.3671
Dirichreg	0.8193	6.2170	15.9382
Beta-Liouville	0.7916	8.7626	7.6380

from two video cameras installed. A centralized database with a web application for continuous data retrieval and monitoring from different sources is included.

The measurements are determined at an interval of half an hour over a period of 15 days. The time series data are considered as individual data points, treated unrelated. Since the output is one-dimensional, the Beta-Liouville regression is reduced to Beta regression. We have compared it with the ordinary least squares linear model regression. The data is normalized, number of data points $n = 720$, dimension $D = 4$. Data are normalized and ensured to be between 0 and 1. Though occupancy estimation is a classification problem, we are employing Beta regression algorithm. This gives us the output in the range of (0-1). It is multiplied with the maximum occupants of the room from test data and rounded off to the nearest integer to get the occupants number. To calculate the fit of regression line, we use the same parameters of comparison as described in the above section.

4.5 Conclusion

In this chapter, we have introduced and investigated a new regression algorithm based on Beta-Liouville distribution, which includes Dirichlet distribution as a special case. The advantage of our method over the well established and widely used technique ordinary-least-squares regression is that it can be viewed as a more general approach with better fit to data. We have illustrated our results with many concrete examples and challenging applications. Namely, share-market analysis of technology based companies with respect to their Google trends, occupancy estimation in an office environment based on inputs from heterogeneous sensors. The proposed regression technique is shown to offer improved results. Further application of our work could focus on applying Beta-Liouville regression to other problems besides market-share analysis and smart buildings, since compositional data are naturally generated in many other research areas, such as Bioinformatics, Chemometrics, image processing and computer vision.

Table 4.2: Beta-Liouville Regression measures of forensic glass data

Methods\Measures	RSS	R2T	KL
CLR + OLS	9.4718	0.0126	216.8152
ILR + OLS	9.4577	0.0129	217.3527
Dirichreg	0.0186	0.0023	333.3006
Beta-Liouville	0.0147	0.00255	306.2127

Table 4.3: Beta-Liouville Regression measures of world-wide browser shares

Methods\Measures	RSS	R2T	KL
CLR + OLS	4.1566	0.9075	0.7468
ILR + OLS	6.2981	0.0386	11.4344
Dirichreg	0.0279	0.8230	0.8385
Beta-Liouville	0.0314	1.0440	0.2551

Table 4.4: Beta-Liouville Regression measures of mobile vendor shares in Canada

Methods\Measures	RSS	R2T	KL
CLR + OLS	4.9768	5.6769	9.6099
ILR + OLS	4.9934	0.3337	9.1590
Dirichreg	0.0004	0.2747	0.0044
Beta-Liouville	0.0441	0.3066	0.0884

Table 4.5: Beta-Liouville Regression measures of Social Media Shares in India

Methods\Measures	RSS	R2T	KL
CLR + OLS	5.0926	0.0419	11.3324
ILR + OLS	5.1004	0.0395	11.4662
Dirichreg	0.6679	0.0741	11.5216
Beta-Liouville	0.1846	0.5477	116.3161

Table 4.6: Regression measures of Smart buildings

Methods\Measures	RSS	R2T	KL
OLS	1.017579	0.8320775	1.330364
Beta-Liouville	0.7270394	0.8205	1.025146

Chapter 5

Conclusion

In this thesis, we have explored in detail different regression techniques for compositional data. We start with different transformations to compositional data, combined with partial least squares discriminant analysis. This is a regression technique used to classify objects. We have applied it to intrusion detection and spam filtering. Further we explore different distribution based regression algorithms. Beta regression and Dirichlet regression are existing approaches for proportional data. We have generalized Beta regression for higher dimensionality. Generalized Dirichlet and Beta-Liouville have been previously used in computer vision applications as a clustering approach. We have exploited these distributions as regression-based techniques.

In order to validate the performance and accuracy of the proposed approach, applications including market-share analysis and occupancy estimation of smart buildings have been conducted and the results are analysed and compared with popular machine learning models. Our study could help to determine if the investment a company makes on advertisements has been fruitful in increasing its market shares by studying how it is trending compared to its competitors in google-trends. Our work could be a useful tool for upcoming companies in the industry to estimate how they are faring in comparison to their competitors in the share market of their product.

Future research directions will focus on model adjustments and improvements to achieve higher regression accuracy. The proposed work could be applied to other applications such as computer vision, Bio-informatics and Chemometrics. Other areas of rerseach could focus on automated model selection, like graphical modelling. All the features available in ”betareg” and ”Dirichreg” packages of R-Studio for beta regression and Dirichlet regression could also be extended to generalized Dirichlet and Beta-Liouville regression algorithms explained in this work. Mixtures of Dirichlet

distributions extended to Dirichlet regression mixture models and mixture models for generalized Dirichlet and Beta-Liouville regression could be implemented.

Bibliography

[1] A. Kalivodová, K. Hron, P. Filzmoser, L. Najdekr, H. Janečková, and T. Adam, "Pls-da for compositional data with application to metabolomics," *Journal of Chemometrics*, vol. 29, no. 1, pp. 21–28, 2015.

[2] N. Bouguila and D. Ziou, "A probabilistic approach for shadows modeling and detection," in *Proceedings of the 2005 International Conference on Image Processing, ICIP 2005, Genoa, Italy, September 11-14, 2005*, 2005, pp. 329–332.

[3] ——, "Unsupervised selection of a finite dirichlet mixture model: An mml-based approach," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 993–1009, 2006. [Online]. Available: https://doi.org/10.1109/TKDE.2006.133

[4] ——, "On fitting finite dirichlet mixture using ECM and MML," in *Pattern Recognition and Data Mining, Third International Conference on Advances in Pattern Recognition, ICAPR 2005, Bath, UK, August 22-25, 2005, Proceedings, Part I*, 2005, pp. 172–182.

[5] ——, "Using unsupervised learning of a finite dirichlet mixture model to improve pattern recognition applications," *Pattern Recognition Letters*, vol. 26, no. 12, pp. 1916–1925, 2005.

[6] N. Bouguila, "Bayesian hybrid generative discriminative learning based on finite liouville mixture models," *Pattern Recognition*, vol. 44, no. 6, pp. 1183–1200, 2011. [Online]. Available: https://doi.org/10.1016/j.patcog.2010.12.010

[7] N. Bouguila and D. Ziou, "Mml-based approach for finite dirichlet mixture estimation and selection," in *Machine Learning and Data Mining in Pattern Recognition, 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005, Proceedings*, 2005, pp. 42–51.

[8] M. Mehdi, E. Epaillard, N. Bouguila, and J. Bentahar, "Modeling and forecasting time series of compositional data: A generalized dirichlet power steady model," in *International Conference on Scalable Uncertainty Management*. Springer, 2015, pp. 170–185.

[9] N. Bouguila and D. Ziou, "A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, p. 107, 2010.

[10] W. Fan and N. Bouguila, "Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference," *IEEE transactions on neural networks and learning systems*, vol. 24, no. 11, pp. 1850–1862, 2013.

[11] N. Bouguila and D. Ziou, "A powerful finite mixture model based on the generalized dirichlet distribution: Unsupervised learning and applications," in *17th International Conference on Pattern Recognition, ICPR 2004, Cambridge, UK, August 23-26, 2004.*, 2004, pp. 280–283. [Online]. Available: https://doi.org/10.1109/ICPR.2004.1334107

[12] ——, "A dirichlet process mixture of dirichlet distributions for classification and prediction," in *2008 IEEE Workshop on Machine Learning for Signal Processing*, 2008, pp. 297–302.

[13] ——, "A hybrid sem algorithm for high-dimensional unsupervised learning using a finite generalized dirichlet mixture," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2657–2668, 2006.

[14] S. Boutemedjet, D. Ziou, and N. Bouguila, "Unsupervised feature selection for accurate recommendation of high-dimensional image data," in *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 2007, pp. 177–184.

[15] N. Bouguila, "Spatial color image databases summarization," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007*, 2007, pp. 953–956.

[16] N. Bouguila and W. ElGuebaly, "A statistical model for histogram refinement," in *International Conference on Artificial Neural Networks*. Springer, 2008, pp. 837–846.

[17] N. Bouguila and D. Ziou, "Dirichlet-based probability model applied to human skin detection [image skin detection]," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 5. IEEE, 2004, pp. V–521.

[18] ——, "Improving content based image retrieval systems using finite multinomial dirichlet mixture," in *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, 2004, pp. 23–32.

[19] N. Bouguila and O. Amayri, "A discrete mixture-based kernel for svms: Application to spam and image categorization," *Information Processing & Management*, vol. 45, no. 6, pp. 631–642, 2009.

[20] J. J. Egozcue and V. Pawlowsky-Glahn, "Groups of parts and their balances in compositional data analysis," *Mathematical Geology*, vol. 37, no. 7, pp. 795–828, 2005.

[21] N. Bouguila, D. Ziou, and R. I. Hammoud, "On bayesian analysis of a finite generalized dirichlet mixture via a metropolis-within-gibbs sampling," *Pattern Analysis and Applications*, vol. 12, no. 2, pp. 151–166, 2009.

[22] S. Boutemedjet, D. Ziou, and N. Bouguila, "Model-based subspace clustering of non-gaussian data," *Neurocomputing*, vol. 73, no. 10-12, pp. 1730–1739, 2010.

[23] J. Aitchison, "The statistical analysis of compositional data," 1986.

[24] M. T. Tsagris, S. Preston, and A. T. Wood, "A data-based power transformation for compositional data," *arXiv preprint arXiv:1106.1451*, 2011.

[25] N. Bouguila, "Count data modeling and classification using finite mixtures of distributions," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 186–198, 2011. [Online]. Available: https://doi.org/10.1109/TNN.2010.2091428

[26] D. Ankam and N. Bouguila, "Compositional data analysis with pls-da and security applications," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2018, pp. 338–345.

[27] M. Pérez-Enciso and M. Tenenhaus, "Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (pls-da) approach," *Human genetics*, vol. 112, no. 5-6, pp. 581–592, 2003.

[28] M. Barker and W. Rayens, "Partial least squares for discrimination," *Journal of chemometrics*, vol. 17, no. 3, pp. 166–173, 2003.

[29] D. R. Perez and G. Narasimhan, "So you think you can pls-da?" *bioRxiv*, p. 207225, 2017.

[30] N. Bouguila and D. Ziou, "A countably infinite mixture model for clustering and feature selection," *Knowl. Inf. Syst.*, vol. 33, no. 2, pp. 351–370, 2012. [Online]. Available: https://doi.org/10.1007/s10115-011-0467-4

[31] M. Andersson, "A comparison of nine pls1 algorithms," *Journal of Chemometrics*, vol. 23, no. 10, pp. 518–529, 2009.

[32] R. G. Brereton and G. R. Lloyd, "Partial least squares discriminant analysis: taking the magic away," *Journal of Chemometrics*, vol. 28, no. 4, pp. 213–225, 2014.

[33] P. Geladi and B. R. Kowalski, "Partial least-squares regression: a tutorial," *Analytica chimica acta*, vol. 185, pp. 1–17, 1986.

[34] G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. J. Egozcue, "The principle of working on coordinates," *Compositional data analysis: Theory and applications*, pp. 31–42, 2011.

[35] H. Lancaster, "The helmert matrices," *The American Mathematical Monthly*, vol. 72, no. 1, pp. 4–12, 1965.

[36] M. Tsagris, S. Preston, and A. T. Wood, "Improved classification for compositional data using the α-transformation," *Journal of Classification*, vol. 33, no. 2, pp. 243–261, 2016.

[37] F. Wentao, N. Bouguila, and D. Ziou, "Unsupervised anomaly intrusion detection via localized bayesian feature selection," in *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, 2011, pp. 1032–1037.

[38] N. Bouguila and D. Ziou, "Online clustering via finite mixtures of dirichlet and minimum message length," *Eng. Appl. of AI*, vol. 19, no. 4, pp. 371–379, 2006.

[39] R. K. Kumar, G. Poonkuzhali, and P. Sudhakar, "Comparative study on email spam classifier using data mining techniques," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2012, pp. 14–16.

[40] N. Bouguila, "Infinite liouville mixture models with application to text and texture categorization," *Pattern Recognition Letters*, vol. 33, no. 2, pp. 103–110, 2012. [Online]. Available: https://doi.org/10.1016/j.patrec.2011.09.037

[41] R. Bace and P. Mell, "Nist special publication on intrusion detection systems," BOOZ-ALLEN AND HAMILTON INC MCLEAN VA, Tech. Rep., 2001.

[42] L. Dhanabal and S. Shantharajah, "A study on nsl-kdd dataset for intrusion detection system based on classification algorithms," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 6, pp. 446–452, 2015.

[43] J. Morais, C. Thomas-Agnan, and M. Simioni, "Using compositional and dirichlet models for market share regression," *Journal of Applied Statistics*, vol. 45, no. 9, pp. 1670–1689, 2018.

[44] R. S. Tay and P. S. Mc Carthy, "Demand oriented policies for improving market share in the us automobile industry," *International Journal of Transport Economics/Rivista internazionale di economia dei trasporti*, pp. 151–166, 1991.

[45] U. Cantner, J. J. Krüger, and R. Söllner, "Product quality, product price, and share dynamics in the german compact car market," *Industrial and Corporate Change*, vol. 21, no. 5, pp. 1085–1115, 2012.

[46] J. Moraisab, C. Thomas-Agnana, and M. Simionic, "A tour of regression models for explaining shares," 2016.

[47] P. Dussauge, B. Garrette, W. Mitchell *et al.*, "The market-share impact of inter-partner learning in alliances: evidence from the global auto industry," *Cooperative strategies and alliances*, pp. 707–727, 2002.

[48] M. J. Maier, "Dirichletreg: Dirichlet regression for compositional data in r," 2014.

[49] Y. Zhang, H. Zhou, J. Zhou, and W. Sun, "Regression models for multivariate count data," *Journal of Computational and Graphical Statistics*, vol. 26, no. 1, pp. 1–13, 2017.

[50] D. Ankam and N. Bouguila, "Generalized dirichlet regression and other compositional models with application to market-share data mining of information technology companies." in

21st International Conference on Enterprise Information Systems (ICEIS). ICEIS, 2019, p. unpublished.

[51] J. Fox and G. Monette, *An R and S-Plus companion to applied regression*. Sage, 2002.

[52] G. D. Hutcheson, "Ordinary least-squares regression," *The SAGE Dictionary of Quantitative Management Research*, pp. 224–228, 2011.

[53] C. L. Bayes, J. L. Bazán, C. García *et al.*, "A new robust regression model for proportions," *Bayesian Analysis*, vol. 7, no. 4, pp. 841–866, 2012.

[54] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, 2004.

[55] R. H. Hijazi and R. W. Jernigan, "Modelling compositional data using dirichlet regression models," *Journal of Applied Probability & Statistics*, vol. 4, no. 1, pp. 77–91, 2009.

[56] N. Bouguila, "On the smoothing of multinomial estimates using liouville mixture models and applications," *Pattern Analysis and Applications*, vol. 16, no. 3, pp. 349–363, 2013.

[57] W.-Y. Chang, R. D. Gupta, and D. S. P. Richards, "Structural properties of the generalized dirichlet distributions," *Contemp. Math*, vol. 516, pp. 109–124, 2010.

[58] C. E. B. Owen, "Parameter estimation for the beta distribution," 2008.

[59] J. Watt, R. Borhani, and A. K. Katsaggelos, *Machine learning refined: foundations, algorithms, and applications*. Cambridge University Press, 2016.

[60] R. H. Hijazi, "Residuals and diagnostics in dirichlet regression," *ASA Proceedings of the General Methodology Section*, pp. 1190–1196, 2006.

[61] N. R. Draper and H. Smith, *Applied regression analysis*. John Wiley & Sons, 2014, vol. 326.

[62] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[63] C. G. Haaf, J. J. Michalek, W. R. Morrow, and Y. Liu, "Sensitivity of vehicle market share predictions to discrete choice model specification," *Journal of Mechanical Design*, vol. 136, no. 12, p. 121402, 2014.

[64] J. A. Martin-Fernandez, M. Bren, C. Barcelo-Vidal, and V. Pawlowsky Glahn, "A measure of difference for compositional data based on measures of divergence," *Proceedings of IAMG*, vol. 99, pp. 211–216, 1999.

[65] M. Tsagris and C. Stewart, "A dirichlet regression model for compositional data with zeros," *Lobachevskii Journal of Mathematics*, vol. 39, no. 3, pp. 398–412, 2018.

[66] M. Tsagris, "Regression analysis with compositional data containing zero values," *arXiv preprint arXiv:1508.01913*, 2015.

[67] H. Choi and H. Varian, "Predicting the present with google trends," *Economic Record*, vol. 88, pp. 2–9, 2012.

[68] S. Vosen and T. Schmidt, "Forecasting private consumption: survey-based indicators vs. google trends," *Journal of Forecasting*, vol. 30, no. 6, pp. 565–578, 2011.

[69] Y. Carrière-Swallow and F. Labbé, "Nowcasting with google trends in an emerging market," *Journal of Forecasting*, vol. 32, no. 4, pp. 289–298, 2013.

[70] T.-T. Wong, "Perfect aggregation of bayesian analysis on compositional data," *Statistical Papers*, vol. 48, no. 2, pp. 265–282, 2007.

[71] F. Cribari-Neto and A. Zeileis, "Beta regression in r," 2009.

[72] A. Gupta and D. Song, "Generalized liouville distribution," *Computers & Mathematics with Applications*, vol. 32, no. 2, pp. 103–109, 1996.

[73] S. Y. Dennis III, "On the hyper-dirichlet type 1 and hyper-liouville distributions," *Communications in Statistics-Theory and Methods*, vol. 20, no. 12, pp. 4069–4081, 1991.

[74] D. Song and A. Gupta, "Properties of generalized liouville disributions," *Random Operators and Stochastic Equations*, vol. 5, no. 4, pp. 337–348, 1997.

[75] A. Sefidpour and N. Bouguila, "Spatial finite non-gaussian mixture for color image segmentation," in *International Conference on Neural Information Processing*. Springer, 2011, pp. 514–521.

[76] M. Amayri, A. Arora, S. Ploix, S. Bandhyopadyay, Q.-D. Ngo, and V. R. Badarla, "Estimating occupancy in heterogeneous sensor environment," *Energy and Buildings*, vol. 129, pp. 46–58, 2016.

[77] D. Ankam, N. Bouguila, and M. Amayri, "Beta-liouville regression and applications." in *6th International Conference on Control, Decision and Information Technologies (CoDIT 2019)*. CoDIT, 2019, p. unpublished.

[78] S. Boutemedjet, N. Bouguila, and D. Ziou, "A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1429–1443, 2009.

[79] N. Bouguila, "Hybrid generative/discriminative approaches for proportional data modeling and classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 12, pp. 2184–2202, 2012.

[80] K. L. Vasconcellos and F. Cribari-Neto, "Improved maximum likelihood estimation in a new class of beta regression models," *Brazilian Journal of Probability and Statistics*, pp. 13–31, 2005.

[81] N. Bouguila and D. Ziou, "Unsupervised selection of a finite dirichlet mixture model: an mml-based approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 993–1009, 2006.

[82] ——, "High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, 2007.

[83] M. Stone and R. J. Brooks, "Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 237–269, 1990.

[84] J. Zhao, B. Lasternas, K. P. Lam, R. Yun, and V. Loftness, "Occupant behavior and schedule modeling for building energy simulation through office appliance power consumption data mining," *Energy and Buildings*, vol. 82, pp. 341–355, 2014.

[85] K. Padmanabh, A. Malikarjuna V, S. Sen, S. P. Katru, A. Kumar, S. K. Vuppala, S. Paul *et al.*, "isense: a wireless sensor network based conference room management system," in *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*. ACM, 2009, pp. 37–42.

[86] J. Page, D. Robinson, and J.-L. Scartezzini, "Stochastic simulation of occupant presence and behaviour in buildings," in *Proc. Tenth Int. IBPSA Conf: Building Simulation*, 2007, pp. 757–764.