Generalized Dirichlet Regression and other Compositional Models with Application to Market-share Data Mining of Information Technology Companies

Divya Ankam[a] and Nizar Bouguila[b]
CIISE, Concordia University, Montreal, Canada

Keywords: Compositional Data, Dirichlet Regression, Generalized Dirichlet, Market-shares, Financial Data Mining.

Abstract: 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.

1 INTRODUCTION

Our aim is to predict the change in market-share (Morais et al., 2018) (Tay and Mc Carthy, 1991) 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 (Cantner et al., 2012). 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 are not obliged to disclose it. 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 (Moraisab et al., 2016) (Dussauge et al., 2002) 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 (Fan and Bouguila, 2013b) (Bouguila and Ziou, 2004b). This will be the prediction.

Mathematically compositional data (Aitchison, 1982) (Fan et al., 2013) are represented in a standard simplex of the sample space given by,

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

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.

Regression problems based on compositional data can be categorized into 2 groups. In the first group the dependant variable is compositional (Ankam and Bouguila, 2018), 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 paper. So far, Dirichlet regression (Maier, 2014) (Bouguila and Ziou, 2004a) has been the best approach in the industry to define compositional regression problems where the dependent variables are compositional.

In this paper, we develop generalized Dirichlet (GD) regression, where the dependent variables follow a generalized Dirichlet distribution (Zhang et al., 2017) (Bouguila and Ziou, 2006). With double the

[a] https://orcid.org/0000-0002-6898-4953
[b] https://orcid.org/0000-0001-7224-7940

Ankam, D. and Bouguila, N.
Generalized Dirichlet Regression and other Compositional Models with Application to Market-share Data Mining of Information Technology Companies.
DOI: 10.5220/0007708201580166
In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 158-166
ISBN: 978-989-758-372-8

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 paper 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 2 describes the machine learning algorithms employed for the research. Section 2.3.3 explains our contribution in devising the GD regression algorithm. Section 3 gives the background for experimental set-up. Section 4 shows our results and analysis. We conclude with Section 5.

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.

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 (Fox and Monette, 2002). 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 (Hutcheson, 2011). If we consider a linear system of variables, with n data points:

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

here β is the regression co-efficient

$$\mathbf{X}\beta = \mathbf{y} \tag{3}$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{4}$$

Such a system usually has no arithmetic solution, rather the aim is to find the best coefficients β which better fit the equations, to solve the quadratic minimization problem

$$\hat{\beta} = \underset{\beta}{\arg\min} S(\beta) \tag{5}$$

where the objective function S is given by

$$S(\beta) = \sum_{i=1}^{m} \left| y_i - \sum_{j=1}^{n} X_{ij}\beta_j \right|^2 = \|\mathbf{y} - \mathbf{X}\beta\|^2 \tag{6}$$

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

$$\hat{\beta} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{7}$$

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 3.2. We discuss two Aitchison transformations (Aitchison, 1982), CLR (centred log ratio transform) and ILR (Isometric log ratio transform). These are widely used in the case of compositional data.

2.2.1 Centered Log Ratio (CLR) + OLS

CLR is defined as:

$$y = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)' \tag{8}$$

CLR can also be represented as follows

$$y = \{y_i\}_{i=1,\ldots,D} = \left\{ \ln \frac{x_i}{g(x)} \right\}_{i=1,\ldots,D} \quad (9)$$

$$g(x) = \left(\prod_{j=1}^{D} x_j \right)^{1/D} \quad (10)$$

Here *g(x)* is the geometric mean of the composition. Resulting in D variables, each representing one component of the original compositional part. Individual contribution of each part is easily interpretable. Subcompositional incoherence and generation of singular matrix are the drawbacks of CLR. Singular data is incompatible with most of the present statistical analysis methods. This can be overcame by ILR transformation.

2.2.2 Isometric Log Ratio (ILR) + OLS

ILR is defined as:

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, i = 1,\ldots,D-1 \quad (11)$$

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

$$z = Hy \quad (12)$$

Where H is the Helmert sub- matrix, obtained by removing the first row of Helmert matrix (Lancaster, 1965). H is a popular orthonormal basis. ILR is applied on data matrix X, subsequent ILR co-ordinates Z are fed into the OLS algorithm in section 2.1. The regression coefficient matrix thus obtained, using equation (3) is γ, has $(D-1)$ x n dimensions. When substituted in equation above, the predictor variables Y are obtained.

2.3 Distributions-based Regression

2.3.1 Beta Regression

Assuming the response data is Beta distributed (Bayes et al., 2012), The authors in (Ferrari and Cribari-Neto, 2004) 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:

$$Y \sim \mathcal{B}(p,q), f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \quad (13)$$

Maximum likelihood estimation is performed to deduce the values of p and q. The closed form solution to this equation is given in (Ferrari and Cribari-Neto, 2004). The partial derivatives of log of Beta distribution with respect to p and q are given by

$$\frac{\partial \log f(y;p,q)}{\partial p} = \psi(p+q) - \psi(p) + \log y \quad (14)$$

$$\frac{\partial \log f(y;p,q)}{\partial q} = \psi(p+q) - \psi(q) + \log(1-y) \quad (15)$$

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

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) dt \quad (16)$$

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

$$E[\log Y] = \psi(p) - \psi(p+q) \quad (17)$$

$$E[\log(1-Y)] = \psi(q) - \psi(p+q) \quad (18)$$

The distribution of response variable $\mathcal{Y}_i$ 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

$$p_i = g(\beta_1 x_{1i} + \cdots + \beta_m x_{mi}) \quad (19)$$

$$q_i = h(\gamma_1 v_{1i} + \cdots + \gamma_M v_{Mi}) \quad (20)$$

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

$$\ell(\theta) = \sum_{i=1}^{n} \log\Gamma(p_i+q_i) - \sum_{i=1}^{n} \log\Gamma(p_i) - \sum_{i=1}^{n} \log\Gamma(q_i) + \sum_{i=1}^{n} p_i \log y_i + \sum_{i=1}^{n} q_i \log(1-y_i) \quad (21)$$

Whose first order derivatives are given by

$$\frac{\partial \ell}{\partial \beta_r} = \sum_{i=1}^{n} g_i' x_{ri} [\psi(p_i+q_i) - \psi(p_i) + \log y_i]$$
$$\frac{\partial \ell}{\partial \gamma_R} = \sum_{i=1}^{n} h_i' v_{Ri} [\psi(p_i+q_i) - \psi(q_i) + \log(1-y_i)] \quad (22)$$

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, the 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.

2.3.2 Dirichlet Regression

Marco J Maier has proposed Dirichlet regression (Maier, 2014) (Hijazi and Jernigan, 2009), which assumes the 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 (Fan and Bouguila, 2013a), defined in equation 23. Also known as common parametrization.

$$\mathcal{D}(y|\alpha) = \frac{1}{\mathrm{B}(\alpha)} \prod_{c=1}^{C} y_c^{(a_c-1)} \tag{23}$$

$$\text{where } B(\alpha) = \prod_{c-1}^{C} \Gamma(\alpha_c) / \Gamma\left(\sum_{c-1}^{C} \alpha_c\right) \tag{24}$$

$$\text{and } \Gamma(x) = \int_0^{\infty} t^{x-1} \exp(-t) \mathrm{d}t \tag{25}$$

Alternately, Dirichlet distribution can also be represented as a function of mean μ and variance ϕ as in equation 26, called alternate parametrization.

$$f(y|\mu,\phi) = \frac{1}{\mathrm{B}(\mu\phi)} \prod_{c=1}^{C} y_c^{(\mu_c\phi-1)} \tag{26}$$

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

$$\ell_c(y|\alpha) = \log\Gamma\left(\sum_{c=1}^{C} \alpha_c\right) - \sum_{c=1}^{C} \log\Gamma(\alpha_c) + \sum_{c=1}^{C} (\alpha_c - 1)\log(y_c) \tag{27}$$

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

$$g(\alpha_c) = \eta_c = X^{[c]}\beta^{[c]} \tag{28}$$

The first order derivative of the log-likelihood:

$$\frac{\partial \ell_c}{\partial \beta_m^{[d]}} = x_m^{[c]} \alpha^{[d]} \left[\log\left(y^{[d]}\right) - \psi\left(\alpha^{[d]}\right) + \psi\left(\sum_{c=1}^{C} \alpha^{[c]}\right)\right] \tag{29}$$

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 (Maier, 2014).

$$\frac{\partial^2 \ell_c}{\partial \beta_m^{[d]} \partial \beta_n^{[d]}} = x_m^{[d]} x_n^{[d]} \alpha^{[d]} \left\{\log(y_d) + \psi\left(\sum_{c=1}^{C} \alpha_c\right) - K\right\} \tag{30}$$

$$K = \psi(\alpha_d) - \alpha_d \left[\psi_1\left(\sum_{c=1}^{C} \alpha_c\right) - \psi_1(\alpha_d)\right] \tag{31}$$

$$\frac{\partial^2 \ell_c}{\partial \beta_m^{[d]} \partial \beta_n^{[e]}} = x_m^{[d]} x_n^{[e]} \alpha_d \alpha_e \psi_1\left(\sum_{c=1}^{C} \alpha_c\right) \tag{32}$$

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.

2.3.3 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 is as follows

$$c \prod_{i=1}^{n} \left[x_i^{a_i-1} \left(1 - \sum_{k=1}^{i} x_k\right)^{b_i-1}\right] \tag{33}$$

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.

$$c = \prod_{i=1}^{n} \frac{\Gamma\left(1 + \sum_{k=i}^{n} (a_k + b_k - 1)\right)}{\Gamma(a_i)\,\Gamma\left(b_i + \sum_{k=i+1}^{n} (a_k + b_k - 1)\right)} \tag{34}$$

In an n dimensional GD distributed data, there are $2n$ variables to estimate. There would be $2n$ log-likelihood equations to solve simultaneously. (Chang et al., 2010) demonstrate that GD can be transformed to n Beta distributions. Suppose $(X_1, \ldots, X_n) \sim GD(a_1, \ldots, a_n; b_1, \ldots, b_n)$ Where $Zi = Zi...Zn$ are n mutually independent Beta distributed variables

$$Z_i \sim B\left(a_i, b_i + \left(\sum_{k=i+1}^{n} (b_k + a_k - 1)\right)\right), i = 1, \ldots, n \tag{35}$$

$$(X_1, \ldots, X_n) \triangleq \left(Z_1, Z_2(1 - Z_1), \ldots, Z_n \prod_{i=1}^{n-1} (1 - Z_i)\right) \tag{36}$$

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

$$Z_i = X_i / \left(1 - \sum_{j=1}^{i-1} X_j\right) \sim B(a_i, c_i) \tag{37}$$

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

$$\prod_{l=1}^{N}\prod_{i=1}^{n}\left[\frac{\Gamma(a_i+c_i)}{\Gamma(a_i)\Gamma(c_i)}\left(\frac{x_{il}}{1-\sum_{j=1}^{i-1}x_{jl}}\right)^{a_i-1}\left(\frac{\sum_{j=1}^{i}x_{jl}}{1-\sum_{j=1}^{i-1}x_{jl}}\right)^{c_i-1}\right] \quad (38)$$

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

$$0=\frac{\partial \log L}{\partial a_i}=\psi(a_i)-\psi(a_i+c_i)-\frac{1}{N}\sum_{l=1}^{N}\log x_{il} \quad (39)$$

$$0=\frac{\partial \log L}{\partial c_i}=\psi(c_i)-\psi(a_i+c_i)-\frac{1}{N}\sum_{l=1}^{N}\log(1-x_{il}) \quad (40)$$

Since a Dirichlet MLE has been transformed to n Beta MLE, we will follow the steps in section 2.3.1 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. The estimated values of regression are normalised to equate to 1 as the expectation is compositional data.

3 EXPERIMENTAL SET-UP

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. Which means, 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 (Watt et al., 2016). 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

procedure K-FOLD-REGRESSION
  for `<s = 1:k>` **do**
    Train a model on sth fold's training set
    Test the model on sth fold's test set
    Compute corresponding evaluation measures
  Compute average of evaluation measures of all k sets

3.2 Evaluation Measures

The goodness of fit of the regression models need 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 (Hijazi, 2006). 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.

3.2.1 R-squared Measure based on Total Variability (R2T)

The term total variability based R square measure was coined by Aitchison in the log ratio analysis[Aitchison 1986]. It is defined as the ratio of total variance in the predicted values to the total variance in the actual values.

$$R_T^2=\text{totvar}(\widehat{\mathbf{x}})/\text{tot}\,var(\mathbf{x}) \quad (41)$$

$$\mathbf{T}(\mathbf{x})=[\tau_{ij}]=\left[\text{var}\left\{\log(x_i/x_j)\right\}\right] \quad (42)$$

$$\text{totvar}(\mathbf{x})=\frac{1}{2d}\sum\mathbf{T}(\mathbf{x}) \quad (43)$$

3.2.2 Residual Sum of Squares (RSS)

Residual sum of squares (RSS) (Draper and Smith, 2014) 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.

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{44}$$

3.2.3 Kullback-Leibler Divergence (KL)

KL divergence is the sum of ratio of the logarithm of actual values o fitted values, weighed by the actual, over each data point(Kullback and Leibler, 1951). Minimum KL divergence is desired as it deduces maximum likelihood (Haaf et al., 2014).

$$KL(\mathbf{S}, \widehat{\mathbf{S}}) = \sum_{t=1}^{T} \sum_{j=1}^{D} \log\left(\frac{S_{jt}}{\widehat{S}_{jt}}\right) S_{jt} \tag{45}$$

KL divergence adapted to compositional data is defined as follows (Martin-Fernandez et al., 1999):

$$KLC(\mathbf{S}, \widehat{\mathbf{S}}) = \frac{D}{2}\left(KL\left(\mathbf{0_D}, \mathbf{S} \ominus \widehat{\mathbf{S}}\right) + KL\left(\mathbf{0_D}, \widehat{\mathbf{S}} \ominus \mathbf{S}\right)\right) \tag{46}$$

$$= \frac{D}{2} \sum_{t=1}^{T} \log\left(\overline{\left(S_t / \widehat{S}_t\right)} \cdot \overline{\left(\widehat{S}_t / S_t\right)}\right) \tag{47}$$

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.

4.1 Real Data

4.1.1 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 (Tsagris and Stewart, 2018) (Tsagris, 2015) and robustness checks (Tsagris, 2015). There are 39 data points with three components. The distribution of the data points is shown in ternary diagram, Figure 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 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.

4.1.2 Glass Composition Dataset

Another dataset containing more number of compositions is the forensic glass dataset (Tsagris and Stewart, 2018). It has 8 components and 214 observations. They are 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 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 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.

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 (Choi and Varian, 2012) (Vosen and Schmidt, 2011). 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 seems 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 effect on the market share. We would further like to explain how the shares have changed over the time, any interesting patterns are ex-

[1] http://gs.statcounter.com

[2] https://trends.google.com/trends/

plained. It is to be noted that google-trends (Carrière-Swallow and Labbé, 2013) 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.

4.2.1 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 2, 3, 4 and 5. This shows how the compositions have changed over the course of 10 years. They clearly follow a linear pattern. Table 3 has 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 transforms combined with OLS with less computational complexity, take lesser time to execute, with acceptable results.

4.2.2 Mobile Seller Market Shares in Canada

The Canadian mobile market was ruled by Apple (88.97%) in the year 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 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.

Table 1: 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
Dirichlet regression	0.8193	6.2170	15.9382
Generalized Dirichlet	0.5415	8.9557	17.9848

Table 2: 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
Dirichlet regression	0.0186	0.0023	333.3006
Generalized Dirichlet	0.0145	0.0015	336.3690

Table 3: 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
Dirichlet regression	0.0279	1.0533	0.2290
Generalized Dirichlet	0.0840	0.8230	0.8385

Table 4: 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
Dirichlet regression	0.0004	0.9443	0.0044
Generalized Dirichlet	0.0100	0.2747	0.1194

Table 5: 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
Dirichlet regression	0.0720	0.0741	11.5216
Generalized Dirichlet	0.6679	0.1262	21.2748

4.2.3 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 5. Dirichlet regression seems to fit the data better than GD regression, with slight variation in the measures.

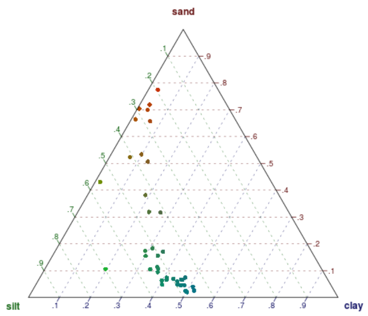

Figure 1: Arctic Lake.

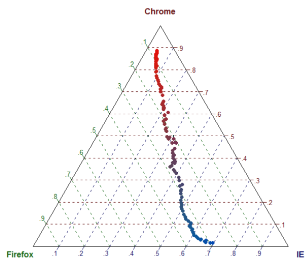

Figure 2: Browser shares: Chrome, IE, Firefox.

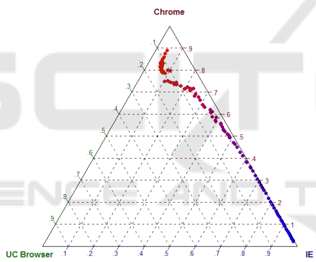

Figure 3: Browser shares: Chrome, IE, UC.

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 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 (Bouguila and Ziou, 2012) in Dirichlet regression. This work could be applied in the fields of image recognition (Boutemedjet et al., 2007) and intrusion detection (Fan et al., 2011).

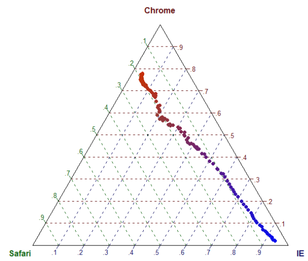

Figure 4: Browser shares: Chrome, Safari, IE.

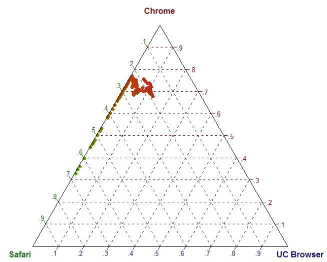

Figure 5: Browser shares: Chrome, UC, Safari.

ACKNOWLEDGEMENTS

The completion of this work was made possible thanks to NSERC and Concordia University research chair.

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 139–177.

Ankam, D. and Bouguila, N. (2018). Compositional data analysis with pls-da and security applications. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 338–345. IEEE.

Bayes, C. L., Bazán, J. L., García, C., et al. (2012). A new robust regression model for proportions. *Bayesian Analysis*, 7(4):841–866.

Bouguila, N. and Ziou, D. (2004a). Dirichlet-based probability model applied to human skin detection [image skin detection]. In *2004 IEEE International Con-*

ference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004, pages 521–524.

Bouguila, N. and Ziou, D. (2004b). 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.*, pages 280–283.

Bouguila, N. and Ziou, D. (2006). A hybrid SEM algorithm for high-dimensional unsupervised learning using a finite generalized dirichlet mixture. *IEEE Trans. Image Processing*, 15(9):2657–2668.

Bouguila, N. and Ziou, D. (2012). A countably infinite mixture model for clustering and feature selection. *Knowl. Inf. Syst.*, 33(2):351–370.

Boutemedjet, S., Ziou, D., and Bouguila, N. (2007). 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*, pages 177–184.

Cantner, U., Krüger, J. J., and Söllner, R. (2012). Product quality, product price, and share dynamics in the german compact car market. *Industrial and Corporate Change*, 21(5):1085–1115.

Carrière-Swallow, Y. and Labbé, F. (2013). Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4):289–298.

Chang, W.-Y., Gupta, R. D., and Richards, D. S. P. (2010). Structural properties of the generalized dirichlet distributions. *Contemp. Math*, 516:109–124.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88:2–9.

Draper, N. R. and Smith, H. (2014). *Applied regression analysis*, volume 326. John Wiley & Sons.

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

Fan, W. and Bouguila, N. (2013a). Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE transactions on neural networks and learning systems*, 24(11):1850–1862.

Fan, W. and Bouguila, N. (2013b). Variational learning of a dirichlet process of generalized dirichlet distributions for simultaneous clustering and feature selection. *Pattern Recognition*, 46(10):2754–2769.

Fan, W., Bouguila, N., and Ziou, D. (2011). 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*, pages 1032–1037.

Fan, W., Bouguila, N., and Ziou, D. (2013). Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference. *IEEE Trans. Knowl. Data Eng.*, 25(7):1670–1685.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Fox, J. and Monette, G. (2002). *An R and S-Plus companion to applied regression*. Sage.

Haaf, C. G., Michalek, J. J., Morrow, W. R., and Liu, Y. (2014). Sensitivity of vehicle market share predictions to discrete choice model specification. *Journal of Mechanical Design*, 136(12):121402.

Hijazi, R. H. (2006). Residuals and diagnostics in dirichlet regression. *ASA Proceedings of the General Methodology Section*, pages 1190–1196.

Hijazi, R. H. and Jernigan, R. W. (2009). Modelling compositional data using dirichlet regression models. *Journal of Applied Probability & Statistics*, 4(1):77–91.

Hutcheson, G. D. (2011). Ordinary least-squares regression. *The SAGE Dictionary of Quantitative Management Research*, pages 224–228.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lancaster, H. (1965). The helmert matrices. *The American Mathematical Monthly*, 72(1):4–12.

Maier, M. J. (2014). Dirichletreg: Dirichlet regression for compositional data in r.

Martin-Fernandez, J. A., Bren, M., Barcelo-Vidal, C., and Pawlowsky Glahn, V. (1999). A measure of difference for compositional data based on measures of divergence. *Proceedings of IAMG*, 99:211–216.

Morais, J., Thomas-Agnan, C., and Simioni, M. (2018). Using compositional and dirichlet models for market share regression. *Journal of Applied Statistics*, 45(9):1670–1689.

Moraisab, J., Thomas-Agnana, C., and Simionic, M. (2016). A tour of regression models for explaining shares.

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

Tsagris, M. (2015). Regression analysis with compositional data containing zero values. *arXiv preprint arXiv:1508.01913*.

Tsagris, M. and Stewart, C. (2018). A dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, 39(3):398–412.

Vosen, S. and Schmidt, T. (2011). Forecasting private consumption: survey-based indicators vs. google trends. *Journal of Forecasting*, 30(6):565–578.

Watt, J., Borhani, R., and Katsaggelos, A. K. (2016). *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge University Press.

Zhang, Y., Zhou, H., Zhou, J., and Sun, W. (2017). Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13.