DISTRIBUTION-BASED REGRESSION FOR COUNT AND SEMI-BOUNDED DATA

Pantea Koochemeshkian

A thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of Master of Applied Science (Electrical and Computer Engineering)
Concordia University
Montréal, Québec, Canada

April 2020

Concordia University
School of Graduate Studies

This is to certify that the thesis prepared

By: **Pantea Koochemeshkian**

Entitled: **Distribution-based Regression for Count and Semi-Bounded Data**

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

Master of Applied Science (Electrical and Computer Engineering)

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

Signed by the final examining commitee:

Dr. Ferhat Khendek _____ Chair and Internal Examiner

Dr. Roch Glitho _____ Examiner

Dr. Joonhee Lee _____ Examiner

Dr. Nizar Bouguila _____ Supervisor

Approved _____
Chair of Department or Graduate Program Director

_____ 2020 _____ _____

Dr. Amir Asif, Dean
Faculty of Engineering and Computer Science

Abstract

Distribution-based Regression for Count and Semi-Bounded Data

Pantea Koochemeshkian

Data mining techniques have been successfully utilized in different applications of significant fields, including pattern recognition, computer vision, medical researches, etc. With the wealth of data generated every day, there is a lack of practical analysis tools to discover hidden relationships and trends. Among all statistical frameworks, regression has been proven to be one of the most strong tools in prediction. The complexity of data that is unfavorable for most models is a considerable challenge in prediction. The ability of a model to perform accurately and efficiently is extremely important. Thus, a model must be selected to fit the data well, such that the learning from previous data is efficient and highly accurate.

This work is motivated by the limited number of regression analysis tools for multivariate count data in the literature. We propose two regression models for count data based on flexible distributions, namely, the multinomial Beta-Liouville and multinomial scaled Dirichlet, and evaluate them in the problem of disease diagnosis. The performance is measured based on the accuracy of the prediction, which depends on the nature and complexity of the dataset. Our results show the efficiency of the two proposed regression models where the prediction performance of both models is competitive to other previously used regression approaches for count data and to the best results in the literature. Then, we propose three regression models for positive vectors based on flexible distributions for semi-bounded data, namely, inverted Dirichlet, inverted generalize Dirichlet, and inverted Beta-Liouville. The efficiency of these models is tested via real-world applications, including software defects prediction, spam filtering, and disease diagnosis. Our results show that the performance of the three proposed regression models is better than other commonly used regression models.

Acknowledgments

I would like to express my sincere appreciation to my supervisor, Professor Nizar Bouguila. As a knowledgeable, respectful and genius supervisor, he always led and motivated me with endless patience. I will be always thankful for his unceasing advices and supports. I was so fortunate to meet him as the instructor of data mining course at Concordia University. There, he introduced the machine learning world to me with his fabulous teaching style. This unique experience was the main motivation to start my research in machine learning with Professor Bouguila's supervision. Being his student will be one of my greatest pleasures in life.

I would like to thanks Dr. Nuha Zamzami for her great support during my work. She always let me have her time whenever I needed help.

Thanks to Narges Manuchehri for her supports and motivations during my challenging time. We worked together on different machine learning algorithms and I learned new topics in this collaboration.

I was fortunate to have such fantastic lab mates, Maryam Rahmanpour, Kamal, Sunny, Omar, Basim, Fatma, Samr, Hussain, Meeta,Muhammad and Hieu who made a memorable time for me.

Last but not least, I am deeply grateful to my parents who encouraged me to work hard, have always supported me and stood by my side to make this achievement. I am heavily indebted to them. I always owe them because of their unconditional love. I would like to express my heartiest thanks to Dr. Hanieh Alipour who supported and motivated me time to time. She is not only my best friend but also my lovely sister in Canada.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Technological advances generate large scale complex data. Thus, retrieval of information and automatically discovering latent patterns have become interesting research topics in various domains of research [1, 2, 3]. Consequently, data mining techniques experienced tremendous development to assist scientists to analyze critical information with minimal human interaction. Data mining techniques have been increasingly attracting the attention of researchers due to their successful application in various fields such as biotechnology, health, microbiology and manufacturing.

Data mining classical techniques can be grouped into three major categories: regression, classification, and clustering. For instance, classification models that describe and distinguish data classes or concepts have been used to analyze information [4]. Classification models are derived based on the analysis of a set of training data where the class labels of the data objects are known, and the model is then used to predict the class labels of unseen objects [5]. Regression [6] has been widely used for prediction on different types of data. It focuses on finding dependencies between objects, and predict target values given training samples of objects and their related target values. This method is called induction [7], and it involves assertions that provide only a finite set of observations. It is commonly recognized that any induction involves some limitations on the presumed dependencies [7].

The majority of the proposed methods make their previous understanding explicit by limiting the range of the presumed dependencies without creating any distributional claims [4, 8]. In this thesis, distribution-based regression approaches using efficient generative models for multivariate discrete data[9] and semi-bounded data

has been proposed.

Regression models have been widely used in the literature as powerful tools to tackle several scientific issues [10]. Examples of successful regression models include multivariate linear regression [11], least-square regression [12], and distribution-based regression for compositional and count data [2, 13, 14]. For instance, [2] has examined regression models for multivariate count data with efficient distributions for analyzing complex genomic data. The authors proposed regression models based on Dirichlet Multinomial and Generalized Dirichlet Multinomial that overcome some limitations of the multinomial model [15]. In this work, we further investigate the problem of analyzing multivariate count responses with other flexible distributions that overcome both specific mean-variance structure and the negative-correlation requirement of the Dirichlet distribution as a prior to the Multinomial. More precisely, two regression models based on Multinomial Beta-Liouville and Multinomial scaled Dirichlet has been proposed.

Several real-life applications naturally generate positive vectors such as visual scenes classification [16]. For instance, in [16], a statistical model based on a finite inverted Dirichlet mixture has been proposed for modeling positive vectors.

Inverted Dirichlet provides good flexibility and simplicity for positive vectors modelling [17] but it has some limitations such as its restrictive strictly positive covariance structure. [16] proposed the Generalized inverted Dirichlet to overcome this problem. [18] proposed a model that is more flexible than the generalized inverted Dirichlet distribution namely the inverted Beta Liouville which contains inverted Dirichlet distribution as a special case.

The mainly focus of the second part of this thesis, is the modeling of positive vectors.

1.1 Related Work

In this section, we review the related works on count data regression. In all the reviewed models here, the dataset symbolized by $\mathcal{X} = \{W_1, \ldots, W_n\}$ which consists of n independent vectors $W_j = (X_j, Y_j)$, where $X_j = (x_{j1}, \ldots, x_{jd})^T$ is a d-dimensional response vector, and $Y_j = (Y_{j1}, \ldots, Y_{jp})^T$ is a p-dimensional co-variate vector.

1.1.1 Dirichlet-Multinomial (DM) Regression

Dirichlet distribution [19], is the multidimensional generalization of the Beta distribution, offering significant flexibility and ease of use. The Dirichlet distribution has the advantage that by varying its parameters [20], it permits multiple modes and asymmetries and can thus approximate a wide variety of shapes [21, 22]. The Dirichlet distribution is commonly used given its flexibility and its several interesting properties, such as the consistency of its estimates, and its ease of use as well as the fact that it is conjugate to the multinomial distribution. Considering the Dirichlet as a prior distribution to the multinomial results in the Dirichlet Multinomial (DM) Distribution [23, 24].

If a d-dimensional count vector $\mathbf{X} = (x_1, \ldots, x_d)$, with $m = \sum_{i=1}^{d} x_i$, follows a multinomial distribution with parameters $\rho = (\rho_1, \ldots, \rho_d)$, then:

$$\mathcal{M}(\mathbf{X}|\rho) = \binom{m}{X} \prod_{i=1}^{d} \rho_i^{x_i} \tag{1}$$

The popular multinomial-logit model uses the joint distribution based on multinomial and Dirichlet [25]. If a vector $\mathbf{X}$ over m possible trails follows the DM Distribution, with parameters $\alpha = (\alpha_1, \ldots, \alpha_d)$, then [2]:

$$\begin{aligned} \mathcal{DM}(\mathbf{X}|\alpha) &= \binom{m}{X} \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + m)} \prod_{i=1}^{d} \frac{\Gamma(x_i + \alpha_i)}{\Gamma(\alpha_i)} \\ &= \binom{m}{X} \frac{\prod_{i=1}^{d} (\alpha_i)_{(x_i)}}{(|\alpha|)_m} \end{aligned} \tag{2}$$

where $(|\alpha|)_{(m)} = |\alpha|(|\alpha| + 1)...(|\alpha| + m - 1)$ denotes the rising factorial, and $|\alpha| = \sum_{i=1}^{d} \alpha_i$.

Even though the DM regression enables the parameterization of the multi-class correlation coefficient for unit-specific covariates, it may disclose additional information that may not be identified by the grouped conditional logit model [26]. The inverse link function $\alpha_i = e^{y^T \alpha_i}$ relates the parameters $\alpha = (\alpha_1, \ldots, \alpha_d)$ of DM distribution to the covariates $\mathbf{X}$. The complete log-likelihood for n independent data

points in this case is given by [2, 26]:

$$\mathcal{L}_n(\mathcal{X}|\alpha) = \sum_{j=1}^{n} ln\binom{m_j}{X_j} + \sum_{i=1}^{d}\sum_{j=1}^{n}\sum_{k=0}^{x_{ij}-1} ln(e^{y_j^T\alpha_i}+k) - \sum_{j=1}^{n}\sum_{k=0}^{m_j-1} ln\left(\sum_{i=1}^{d} e^{y_j^T\alpha_i}+k\right) \tag{3}$$

Estimating the Dirichlet multinomial regression model does not present any specific challenge, and its numerical optimization process based on the Newton-Raphson algorithm provides quick convergence to the maximum [26]. However, the Dirichlet has some disadvantages, such as its very restrictive negative covariance matrix and the fact that the variables with the same mean must have the same variance, which limits its applicability to many data sets [27, 28]. To handle these disadvantages, [2], proposed a regression model with a more flexible mean-covariance and correlation structure based on the generalized Dirichlet multinomial distribution [27].

1.1.2 Generalized Dirichlet Multinomial (GDM) Regression

The Generalized Dirichlet (GD) distribution was introduced in [29], and it has a more general covariance structure than the Dirichlet distribution. The generalized Dirichlet distribution, in fact, can release both constraints of the Dirichlet distribution, including the negative-correlation and the equal-confidence requirements. Thus, it has shown to be a more appropriate prior in Bayesian learning situations [27, 30]. Similar to the Dirichlet, the generalized Dirichlet is a conjugate to the multinomial distribution, but it is more practical for several real-life applications [27, 31]. The composition of the generalized Dirichlet and the multinomial gives the Generalized Dirichlet Multinomial (GDM) distribution. The probability mass of a GDM for a count vector $X = (x_1, \ldots, x_d)$ with a parameter set $\xi = (\alpha_1, \ldots, \alpha_{d-1}, \beta_1, \ldots, \beta_{d-1})$ and $\alpha_i, \beta_i > 0$, is given by [2, 27]:

$$\begin{aligned}\mathcal{GDM}(\mathbf{X}|\xi) &= \binom{m}{X}\prod_{i=1}^{d-1}\frac{\Gamma(\alpha_i+x_i)}{\Gamma(\alpha_i)}\frac{\Gamma(\beta_i+z_{i+1})}{\Gamma(\beta_i)}\frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i+\beta_i+z_i)} \\ &= \binom{m}{X}\prod_{i=1}^{d-1}\frac{(\alpha_i)_{x_i}(\beta_i)_{z_{i+1}}}{(\alpha_i+\beta_i)_{z_i}}\end{aligned} \tag{4}$$

where $z_i = \sum_{l=i}^{d} x_l$ is the cumulative sum.

For relating the covariates $\mathbf{X}$ to the parameters, the following link functions have been used by [2]: $\alpha_i = e^{y^T \alpha_i}$, and $\beta_i = e^{y^T \beta_i}$. Now, let the parameter set $\xi = \{\alpha, \beta\}$ represents all the regression coefficients, the log-likelihood is given by [2]:

$$\mathcal{L}_n(\mathcal{X}|\xi) = \sum_{j=1}^{n} ln\binom{m_j}{X_j} + \sum_{i=1}^{d}\sum_{j=1}^{n}\left(\sum_{k=0}^{x_{ij}-1} ln(e^{y_j^T \alpha_i} + k) + \sum_{k=0}^{z_{i,j+1}-1} ln(e^{y_j^T \beta_i}) - \sum_{k=0}^{z_i-1} ln(e^{y_j^T \alpha_i} + e^{y_j^T \beta_i} + k)\right) \tag{5}$$

Indeed, Generalized Dirichlet Multinomial is a more suitable distribution for modeling count data than the widely used Dirichlet Multinomial. It acquires its flexibility from the fact that Generalized Dirichlet has a more flexible covariance structure, and it has one more set of parameters that grants it a $d-1$ extra degrees of freedom to better fit real data. Both DM and GDM have been well studied in the literature (see, for instance, [2, 26, 27] for more details). In this thesis, the two novel regression models based on alternative distributions that have shown superior performance in modeling count data, namely; Multinomial Beta-Liouville (MBL) distribution [32], and Multinomial scaled Dirichlet (MSD) distribution [33, 34] has been introduced.

1.2 Contribution

Our major contributions in this thesis are as follows:

1. We propose novel regression models for multivariate count data. Our proposed framework is based on Multinomial scaled Drichlet and multinomial beta-liouville distributions. Developed all the equations related to its parameters estimation.This work has been accepted by journal of Cybernetics and Systems [35].

2. We propose novel regression models for semi-bounded data. Our proposed models are based on inverted Dirichlet, generalized inverted Dirichlet and inverted beta-liouville distributions. This work has been submitted to IEEE SMC conference [36].

3.Comparing our models with other related state of the art approaches.

4.Investigation of the performance of our framework by testing it on real data sets as well as real-life applications such as disease diagnosis, spam detection, software modules defect prediction and age prediction.

1.3 Thesis Overview

The rest of this thesis is organized as follows:

- In chapter 2, the Multinomial beta-liouville regression and Multinomial scaled Dirichlet regression models and show the results of our proposed models on real applications has been proposed.
- In chapter 3, three regression models for semi-bounded data applied to real applications such as software defect detection, spam filtering, and age prediction has been propose.
- In chapter 4, conclusion and briefly summarize the contributions and recommend future works.

Chapter 2

Flexible distribution-based regression models for count data: application to medical diagnosis

We propose distribution-based regression approaches using efficient generative models for multivariate count data. Moreover, we investigate the problem of analyzing multivariate count responses with other flexible distributions that overcome both specific mean-variance structure and the negative-correlation requirement of the Dirichlet distribution as a prior to the Multinomial. More precisely, two regression models based on flexible distributions for count data, namely; Multinomial Beta-Liouville and Multinomial scaled Dirichlet has been proposed. First, the response distributions has been introduced, propose the link functions, and derive the score and information matrices for estimating the parameters and give the complete regression algorithm. Furthermore, we investigate, with the proposed models, the problem of the diagnosis of three different diseases, namely, heart attack, breast cancer, diabetes, as well as the analysis of genomics dataset.

The rest of this chapter is organized as follows. In Section 2.1, propose two distribution-based regression models where first to discuss the properties of the considered distributions, then propose the link functions and provide all the details about the models' parameters estimation. Section 2.2 is devoted to the application of the proposed models on real genomics and medical data and to the discussion of the results.

2.1 The proposed regression models

In this section, the details of the proposed models for multivariate count responses has given. For each proposed model, first discuss the properties of the fitting distribution, then proposed the link functions and discuss the maximum likelihood estimation procedure. Finally, the complete learning algorithm has been given.

2.1.1 The Considered Distributions

The Multinomial Beta-Liouville (MBL) distribution

The Liouville family[37] of the second kind includes the Dirichlet distribution as a special case if all variables in the Liouville random vector have the same normalized variance, and the density generator variate has a Beta distribution [28]. Choosing the Beta distribution as a generating density results in which is commonly called the Beta-Liouville distribution [38]. Like the Dirichlet, the Beta-Liouville is a conjugate prior to the multinomial distribution, and it can overcome the main restrictions of the Dirichlet distribution. Moreover, the two more parameters in Beta-Liouville can be used to adjust the spread of the distribution, which makes it more practical and provides better modeling capabilities. Considering the Beta-Liouville as a prior to the multinomial results in a flexible joint distribution called the Multinomial Beta-Liouville (MBL) [32].

The probability of a count vector $\mathbf{X} = (x_1, \ldots, x_d)$ over $m = \sum_{i=1}^{d} x_i$ trials following the MBL model with a parameters set $\theta = (\alpha_1, \ldots, \alpha_{d-1}, \alpha, \beta)$, is given by [32]:

$$\begin{aligned} \mathcal{MBL}(\mathbf{X}|\theta) &= \binom{m}{X} \frac{\Gamma(\sum_{i=1}^{d-1} \alpha_i)\Gamma(\alpha+\beta)\Gamma(\alpha')\Gamma(\beta') \prod_{i=1}^{d-1} \Gamma(\alpha_i')}{\Gamma(\sum_{i=1}^{d-1} \alpha_i')\Gamma(\alpha'+\beta')\Gamma(\alpha)\Gamma(\beta) \prod_{i=1}^{d-1} \Gamma(\alpha_i)} \\ &= \binom{m}{X} \frac{(\alpha)_{z_i} + (\beta)_{x_{i+1}} + (\alpha_i)_{x_i}}{|\alpha|_m (\alpha+\beta)_{z_i}} \end{aligned} \tag{6}$$

where $(a)_{(k)} = a(a+1)..(a+k_1)$ [1], $z_i = \sum_{k=1}^{i} x_k$, $\alpha_i' = \alpha_i + x_i$, $\alpha' = \alpha + \sum_{i=1}^{d} x_i$ and $\beta' = \beta + x_d$. Note that when $\alpha = \sum_{i=1}^{d-1} \alpha_i$ and $\beta = \alpha_d$, the MBL is reduced to the Dirichlet Multinomial (Eq. 2). Indeed, MBL is an attractive distribution to fit count data, given the fact that it has fewer parameters than MGD with a comparable performance [32].

Multinomial scaled Dirichlet (MSD) distribution

The scaled Dirichlet [39] is another generalization of the Dirichlet distribution, which has been proposed to overcome the Dirichlet limitation of not considering the similar positions between categories or multinomial cells. Besides, it has a general and more flexible variance and covariance structure, given the fact that it has one more parameter to model the variance of each dimension independently. Furthermore, the scaled Dirichlet has shown to be an interesting prior to the multinomial, resulting in an efficient hierarchical Bayesian model called Multinomial Scaled Dirichlet (MSD) proposed by [33]. Indeed, MSD has shown to have high flexibility in count data modeling with superior performance in many challenging applications [33, 34, 40].

The scaled Dirichlet has two parameters such that $\alpha = (\alpha_1, \ldots, \alpha_d)$ is the shape parameters and $\beta = (\beta_1, \ldots, \beta_d)$ is the scale parameter [39]. It is noteworthy that when all elements of vector β are equal to some constant, the scaled Dirichlet distribution is reduced to the Dirichlet. Therefore, the scaled Dirichlet with d extra parameters is more flexible than the Dirichlet distribution [41, 42, 43]. If a count vector $\mathbf{X} = (x_1, \ldots, x_d)$, and $m = \sum_{i=1}^{d} x_i$, follows a multinomial scaled Dirichlet, with a set of parameters $\vartheta = \{\alpha, \beta\}$ and $|\alpha| = \sum_{i=1}^{d} \alpha_i$, then [33]:

$$\begin{aligned} \mathcal{MSD}(\mathbf{X}|\vartheta) &= \binom{m}{X} \frac{\Gamma(|\alpha|)}{\Gamma(m+|\alpha|) \prod_{i=1}^{d} \beta_i^{x_i}} \left[\prod_{i=1}^{d} \frac{\Gamma(x_i+\alpha_i)}{\Gamma(\alpha_i)} \right] \\ &= \binom{m}{X} \frac{\prod_{i=1}^{d} (\alpha_i)_{(x_i)}}{(|\alpha|)_m \prod \beta_i^{x_i} \alpha_i} \end{aligned} \tag{7}$$

2.1.2 The proposed link functions

The link function [44] can be defined as the inverse of cumulative distribution function of a continuous distribution. This is used to associate the regression parameters to covariates. Such function provides the relation between the linear prediction and the mean of the distribution function [45]. When considering a distribution function with the canonical parameter, there is always a well-defined canonical link function obtained from the exponential density function of the response [46, 47].

Proposed link functions for MBL regression

For Multinomial Beta-Liouville distribution-based regression, the relation between the parameters and the p-dimensional co-variate vector $X = (x_1, \ldots, x_p)$, has been written in the following forms:

$$\begin{aligned} \alpha_i &= g_1(\alpha_i x_1 + \alpha_i x_2 + \ldots + \alpha_i x_p), \qquad i = 1, \ldots, d \\ \alpha &= g_2(\alpha x_1 + \alpha x_2 + \ldots + \alpha x_p), \\ \beta &= g_3(\beta x_1 + \beta x_2 + \ldots + \beta x_p) \end{aligned} \tag{8}$$

For finding $g(\mu_j)$, the following procedure to be considered:

$$g(\mu_j) = X_j^T \theta \qquad j = 1, \ldots, n \tag{9}$$

where μ_j is the mean of X_j, and θ is a vector of regression parameters. Thus:

$$logit(\mu_j) = \log\Big(\frac{\mu_j}{1-\mu_j}\Big), \tag{10}$$

and for *logit* link function we have the following :

$$\Pi_j(x) = \frac{\exp(\theta^T X_j)}{1 + \sum_{j=1}^{n-1} \exp(\theta^T X_j)} \tag{11}$$

Thus, The following equations for the Multinomial Beta-Liouville model:

$$\begin{aligned} g_1(\mu_j) &= X_j^T \alpha_i \\ g_2(\mu_j) &= X_j^T \alpha \\ g_3(\mu_j) &= X_j^T \beta \end{aligned} \tag{12}$$

The final regression equation as a linear regression equation has been considered:

$$Y = \eta_0 + \eta_1 x_1 + \ldots + \eta_i x_d \tag{13}$$

where $\eta_i = \beta\alpha\alpha_i, i = 1, \ldots, d$, and d is the dimension of the response vector.

Consider the parameters set $\theta = (\alpha_1, \ldots, \alpha_{d-1}, \alpha, \beta)$ as all the regression coefficients, the complete log-likelihood is given by:

$$\begin{aligned}\mathcal{L}_n(\mathcal{X}|\theta) = \sum_{j=1}^{n} ln\binom{m_j}{X_j} + \sum_{i=1}^{d}\sum_{j=1}^{n}\Bigg[\sum_{k=0}^{x_{ij}-1} ln(e^{x_j^T\alpha_i} + k) \\ + \sum_{k=0}^{z_{ij}} ln(e^{x_j^T\beta} + k) + \sum_{k=0}^{x_{i-1}} ln(e^{x_j^T\alpha} + k) \\ - \sum_{k=0}^{x_{i,m}-1} ln(e^{x_j^T\alpha_i} + k) - \sum_{k=0}^{x_{i+1}} ln(e^{x_j^T\alpha} + e^{x_j^T\beta} + k)\Bigg]\end{aligned} \tag{14}$$

Proposed link functions for MSD regression

For multinomial scaled Dirichlet, we can link the parameter $\vartheta = \{\alpha, \beta\}$ to the p-dimensional covariates vector X, as:

$$\alpha_i = \lambda_1(\alpha_i x_1 + \alpha_i x_2 + \ldots + \alpha_i x_p) \tag{15}$$

$$\beta_i = \lambda_2(\beta_i x_1 + \beta_i x_2 + \ldots + \beta_i x_p), \qquad i = 1, \ldots, d \tag{16}$$

For finding the $\lambda(\mu_j)$ following procedure has been followed:

$$\lambda(\mu_j) = X_j^T\vartheta, \qquad j = 1, \ldots, n \tag{17}$$

then we have:

$$\lambda_1(\mu_j) = X_j^T\alpha_i \tag{18}$$

$$\lambda_2(\mu_j) = X_j^T\beta_i \tag{19}$$

Considering the final regression equation to be similar to the linear regression equation, as previously mentioned in Eq.(13), where η_i in case of MSD model is given by $\eta_i = \beta_i\alpha_i, i = 1, \ldots, d$. The complete log-likelihood of MSD for n independent data points is, thus, computed as follows:

$$\begin{aligned}\mathcal{L}_n(\mathcal{X}|\vartheta) = \sum_{j=1}^{n} ln\binom{m_j}{X_j} - \sum_{j=1}^{n}\sum_{k=0}^{x_i}\sum_{i=1}^{d} x_i ln(e^{x_j^T\beta_i} + k) \\ + \sum_{j=1}^{n}\sum_{k=0}^{x_i}\sum_{i=1}^{d}\Big(ln(x_i + e^{x_j^T\alpha_i} + k) - ln(e^{x_j^T\alpha_i + k})\Big) \\ + \sum_{j=1}^{n}\sum_{i=1}^{d}\sum_{k=1}^{x_i}\Big(ln(|e^{x_j^T\alpha_i} + k|) - ln(m_j + |e^{x_j^T\alpha_i} + k|)\Big)\end{aligned} \tag{20}$$

2.1.3 Parameters Estimation

For estimating the parameters, to find the best coefficients for our regression models, the Maximum Likelihood Estimation (MLE) technique [48] was utilized. Maximum likelihood estimation [49, 50], is a method that attempts to discover the most probable model that generated the observed result. The maximum likelihood parameter estimates can obtain for Multinomial Beta-Liouville and Multinomial scaled Dirichlet models by taking the derivative of the complete log-likelihood function, and find Θ when the derivative is equal to zero. In this technique, the estimation of the parameters that maximize the log-likelihood is based on the following:

$$\Theta^{(t+1)} = \arg\max_{\Theta} \sum_{j=1}^{n} \log(p(X_j|\Theta)) \tag{21}$$

For both models, closed-form solutions do not exist. Thus, the process requires a Newton-Raphson optimization that iterates between scoring steps based on the present values and an update of the parameters, such that:

$$\Theta^{(t+1)} = \Theta^{(t)} - H_{\Theta}^{-1} G_{\Theta} \tag{22}$$

where G is the gradients and H is the Hessian matrix based on the first and second order derivatives of the log-likelihood function, respectively. The complete derivations needed for estimating the parameters of the two proposed models are given as follows.

2.1.4 MLE for the proposed models

1. The derivatives to estimate the MBL-based model parameters

The first derivatives of MBL log-likelihood with respect to the regression coefficients are given by:

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_i} = \sum_{j=1}^{n} g_1'(x_j)[\,\psi(\alpha_i) + \psi(\alpha_i') - \psi(\alpha_i') - \psi(\alpha_i)] \tag{23}$$

where $\alpha_i' = \alpha + x_i$.

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha} = \sum_{j=1}^{n} g_2'(x_j)[\,\psi(\alpha + \beta) + \psi(\alpha') - \psi(\alpha' + \beta') - \psi(\alpha)] \tag{24}$$

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \beta} = \sum_{j=1}^{n} g_3'(x_j)[\,\psi(\alpha + \beta) + \psi(\beta') - \psi(\alpha' + \beta') - \psi(\beta)] \tag{25}$$

where $\alpha' = \alpha + \sum_{i=1}^{d} x_i$ and $\beta' = \beta + x_i$. According to Newton-Raphson method, the second-order derivatives should calculate as follows:

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_{i1} \partial \alpha_{i2}} = \sum_{j=1}^{n} g_1''(x_j)[\,\psi'(\alpha_i + \psi'(\alpha_i') - \psi'(\alpha_i' - \psi'(\alpha_i)] \tag{26}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \alpha} = \sum_{j=1}^{n} g_2''(x_j)[\,\psi'(\alpha + \beta) + \psi'(\alpha') - \psi'(\alpha' + \beta') - \psi'(\alpha)] \tag{27}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \beta} = \sum_{j=1}^{n} g_3''(x_j)[\,\psi'(\alpha + \beta) + \psi'(\beta') - \psi'(\alpha' + \beta') - \psi'(\beta)] \tag{28}$$

2. The derivatives to estimate the MSD-based model parameters

The first derivatives of MSD log likelihood function with respect to $\alpha_i, i = 1, \ldots, d$ and $\beta_i, i = 1, \ldots, d$ are given by:

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\vartheta)}{\partial \alpha_i} = \sum_{j=1}^{n} \hat{\lambda}_1(x_j) \left(\Psi(|\alpha|) - \Psi(m_i + |\alpha|) + \Psi(x_i + \alpha_i) - \Psi(\alpha_i)\right) \tag{29}$$

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\vartheta)}{\partial \beta_i} = \sum_{j=1}^{n} \hat{\lambda}_2(x_j) \left(\frac{\mathbf{x}_i}{\beta_i}\right) \tag{30}$$

By computing the second derivatives with respect to α_i and β_i, obtained:

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\vartheta)}{\partial \alpha_{i1} \alpha_{i2}} = \begin{cases} \sum_{j=1}^{n} \hat{\lambda}_1(x_j) \left[\Psi'(|\alpha|) - \Psi'(m_j + |\alpha|) + \Psi'(x_i + \alpha_i) - \Psi'(\alpha_i)\right] \\ \text{if} \quad i_1 = i_2 = i \\ \sum_{j=1}^{n} \hat{\lambda}_1(x_j)[\Psi'(Z) - \Psi'(m_j + |\alpha|)] \quad \text{otherwise,} \end{cases} \tag{31}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\vartheta)}{\partial \beta_{i1} \beta_{i2}} = \begin{cases} \sum_{j=1}^{n} \hat{\lambda}_2(x_j) \left(-\frac{\mathbf{x}_i}{\beta_i^2}\right) & \text{if} \quad i_1 = i_2 = i, \\ 0 & \text{otherwise,} \end{cases} \tag{32}$$

To achieve an optimal performance of our proposed models, the initial values of the parameters were calculated using the method of moments [51], which depends on the mean and variance of each distribution. Then, using the maximum likelihood approach, the parameters are updated to get their natural values with respect to the given dataset. Finally, the regression model is applied to predict the multivariate count response. The complete learning algorithm is summarized in (Algorithm 1).

Algorithm 1 The complete learning algorithm for predicting multivariate count response.

1. **Input** DATA SET $\mathcal{X} = \{W_1, \ldots, W_n\}$ with n independent data points $W_j = (X_j, Y_j)$, where X_j is the count response vector and Y_j is covariate vector.
2. **Output** The final parameters Θ, log-likelihood, predicted Y
3. Split the data by ratio 60:40 for training and testing
4. Initialize the parameters for each model $\Theta^{(0)}$
5. **repeat**
6. Update the parameters $\Theta^{(t)}$ using Eq.(66)
7. Update the link functions
8. Calculate the log-likelihood using Eq.(14) for MBL or Eq.(20) for MSD
9. **until** *convergence*
10. Predict the covariate values of Y using Eq. (13).

2.2 Experimental Results

Our aim in this section is to apply the proposed regression models on real datasets. Multinomial Beta-Liouville and Multinomial scaled Dirichlet regression models has been evaluated to show their effectiveness compared to the previously proposed distribution based regression models for count data.

2.2.1 Data and Performance Measures

The evaluation of each model is based on the Akaike information criterion (AIC) [52], Bayesian information criterion (BIC) [53], and MSE where the smaller values for AIC, BIC and MSE indicate that the model has a better performance. Furthermore, to considered the prediction accuracy where the higher accuracy indicate the better performance of the model. The considered performance metrics are defined as follows:

- **Akaike Information Criterion (AIC)**: AIC is a way of measuring that can be used to assess the capabilities of the model by showing a link between Kullback-Leibler information [54], and maximized log-likelihood [52]. It selects the model that minimises the mean squared or prediction error [55]. The AIC of each model can calculate by using the following formula where N_X is the number of data points:

$$AIC = -2\mathcal{L}_n + 2N_X \tag{33}$$

- **Bayesian Information Criterion (BIC)**: BIC can be extracted from a large-sample approximation [53]. BIC criterion selects the model with the smallest value. For each model, the following is used formula to calculate it, where N_X is the number of data points and D_X is the dimension of the data:

$$BIC = -2\mathcal{L}_n + N_X \log(D_X) \tag{34}$$

- **Accuracy**: Our goal is to predict precision covariate values of Y, which consists of one or more positive values. To find the accuracy of the prediction, $Y_{predict}$ compared to the actual data in the test split Y_{Test} of a given dataset. Since The data is multivariate, where each Y is a vector, the average accuracy was calculated. That is, the average of the differences between $Y_{predict} = (y'_1, \ldots, y'_p)$ and $Y_{Test} = (y_1, \ldots, y_p)$ should be calculated. The following equation is used to calculate the accuracy for each model:

$$ACC = \Big(1 - \frac{\mu\big(|Y_{predict} - Y_{Test}|\big)}{\mu\big(|Y_{Test}|\big)}\Big) \times 100 \tag{35}$$

- **MSE cost function**: The root mean square error (RMSE) is used to measure the performance of the models [56]. This metric majorly presume $i = 1, 2, ...n$ samples of model errors. RMSE formula is given as follows:

$$\mathcal{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \tag{36}$$

Therefore, to train a regression model, it is necessary to find the value of regression coefficients that minimize the RMSE.In practice, it is simpler to minimize the Mean Square Error (MSE). Because the value that minimizes a function

also minimizes its square root, instead of minimizing the RMSE, the regression coefficients that minimize the MSE can be find as given :

$$\mathcal{MSE}(\mathcal{X}, \theta) = \frac{1}{D} \sum_{i=1}^{D} (Y_{predict} - Y)^2 \tag{37}$$

2.2.2 Real Data

The models have been applied to four different applications from the medical domain research field as following:

- Analysis of genomics data: RNA-seq [2].
- Impact of stress on heart attack [57].
- Breast Cancer diagnosis [58].
- Diabetes diagnosis [59].

The evaluation of each model is based on four metrics: Akaike information criterion (AIC) [52], Bayesian information criterion (BIC) [53], log-likelihood and accuracy. The prediction results in the following subsection are shown by different figures, and in each figure the X_axis shows the observed data points, and Y_axis shows the value of each Y that is the prediction value.

2.2.3 Analyzing Genomics Data: RNA-seq

In this application, the problem of high-throughput data analysis in genomics has been studied. Quantifying the genomic features depends on sequencing technology, where the data obtained from sequencing technologies are often summarized by the counts of DNA or RNA fragments within a genomic interval. The RNA-seq (RS) dataset [1] [60] is considered. The data consists of six exons that present the gene, and these six exons in our regression model are exploratory variables where each observation has the expression level with four covariates: total reads, treatment, gender, and age. The total number of observations is 200. Table 1 presents the results of the four tested models, where compared based on AIC, BIC, and accuracy. As the Table 1 shown, the MSD based model has the smallest AIC, BIC, and the highest likelihood.

[1] https://github.com/Yiwen-Zhang/MGLM/tree/master/MGLM/data

In terms of accuracy, the MBL based model outperforms all the tested models with an accuracy of 98% compared to 94-95% for the other models.

Table 1: Models performance comparison for RNA-seq dataset.

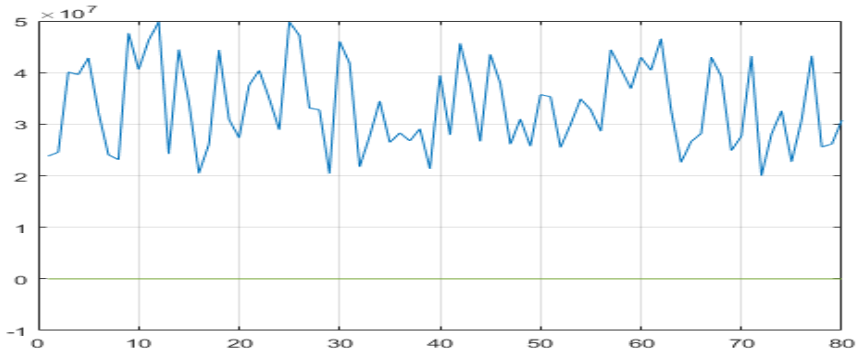
Model	Performance metrics			Accuracy
	Log-likelihood	AIC	BIC	
DM	-1.2634e+03	2.5748e+03	2.6417e+03	94.00%
GDM	-1.1432e+03	2.8721e+03	2.8617e+03	95.00%
MBL	-7.0738e+19	1.4148e+20	1.4148e+20	**98.00%**
MSD	9.5334e+04	**-1.9045e+05**	**-1.9043e+05**	95.75%

Figure 1, and Figure 2 show the predicted values $\hat{Y}$, *i.e.* the values of each of the four attributes (total reads, treatment, gender, and age) that predicted for each observation using MBL and MSD based regression models, respectively. From these figures, can see that the prediction of Y has the same behavior of Y_Test. Note that the small predicted values are approximated to zero. In general, we can say that the predicted values are approximately similar to the actual test values, as shown in the figures, the MBL-based regression model have an accuracy of 98% when using , and 95.75% using the regression model based on MSD.

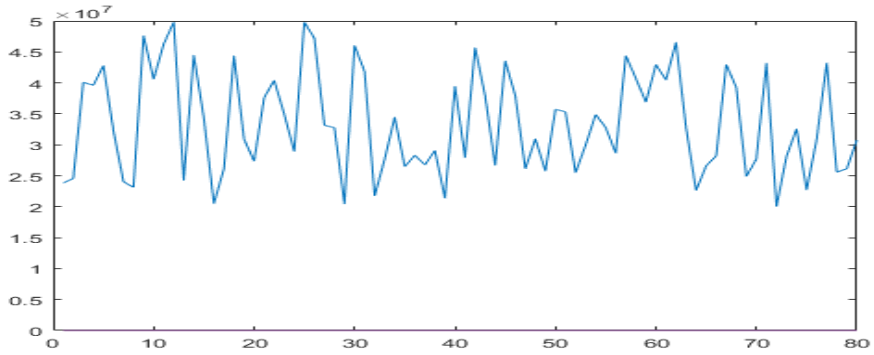
2.2.4 Predicting Heart Attack Risk

This application is based on a publicly available dataset named as Stress Echocardiography (SEG) [2]. The dataset represents a study that has been done to determine the impact of the dobutamine drug on having a risk of heart attack or cardiac event. The observations of the dataset were based on a test that the patient should take through raising the patient's heart rate by the run on the treadmill and gather the needed information. In our experiments, we focus on predicting the cardiac death, *i.e.* the risk of heart attack, to identify predictors of subsequent cardiac events from clinical and demographic information for each patient. Independent variables evaluated were: history of hypertension, diabetes mellitus, MI, CABG, or PTCA, age, gender, peak dose of dobutamine, rest and peak dobutamine heart rate, blood pressure, and rate pressure product (RPP), percent of achieved maximum predicted heart rate, rest and peak dobutamine EF, presence of induced chest pain, negative, equivocal or ischemic

[2]http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/stressEcho.html



(a) Y_Train

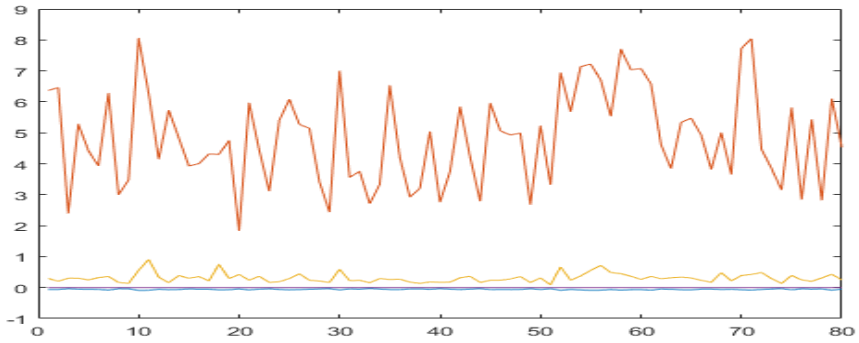


(b) Y_Test

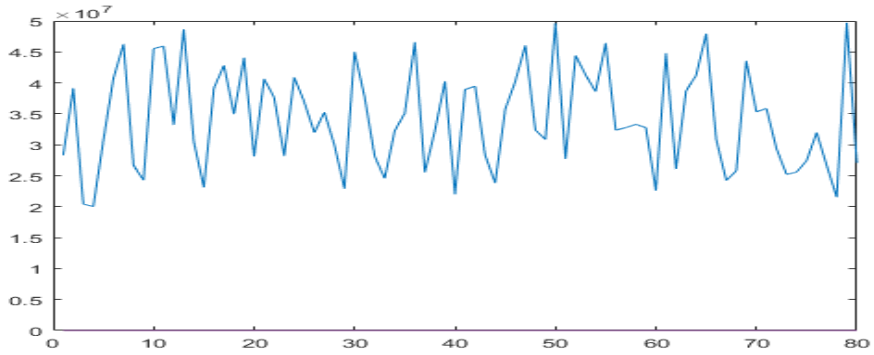
Figure 1: Comparison of test values and the predicted values of Y using MBL-based regression model for RNA-seq dataset.

electrocardiogram (ECG), rest wall-motion abnormality (WMA), and a positive stress echocardiogram (SE) [57]. Then, the prediction of the cardiac events that aimed to predict broken down into four categories (values), representing myocardial infarction (MI), revascularization by percutaneous transluminal coronary angioplasty (PTCA), coronary artery bypass grafting surgery (CABG), and cardiac death.

The prediction results for the considered dataset using the four tested regression models are given in Table 2 reported using the above-mentioned performance metrics. According to the results, one may notice that the DM-based regression model has the smallest likelihood and lowest accuracy as compared to the other tested models. On the other hand, GDM, MBL, and MSD have approximately similar performance according to the prediction accuracy, yet, MSD has a larger log-likelihood and smaller AIC and BIC. Thus, MSD based regression has the best prediction results on this dataset.



(a) Y_Train



(b) Y_Test

Figure 2: Comparison of test values and the predicted values of Y using MSD-based regression model for RNA-seq dataset.

In Figure 3, and Figure 4 displayed the predicted four values of Y , *i.e.* MI, PTCA, CABG and cardiac death, corresponding to each observation in Echocardiography dataset using MBL and MSD based regression models, respectively. From these figures can conclude that the MBL-based regression model has the highest achieved accuracy of 98.90%, which is slightly better for large numbers but not suitable for predicting small values. Furthermore, Figure 4 illustrates that the MSD-based regression model performs well on both large and small values.

2.2.5 Breast Cancer Diagnosis

In this application, Breast Cancer Wisconsin dataset (BCD)[3]has been used, which has a total of 569 observations, and each observation is computed from a digitized

[3]https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Table 2: Models performance comparison for Stress Echocardiography dataset.

Model	Performance metrics			Accuracy
	Log-likelihood	AIC	BIC	
DM	-6.1365e+03	1.2333e+04	1.2447e+04	95.00%
GDM	-5.7684e+03	1.1633e+04	1.1816e+04	98.00%
MBL	-5.6532e+06	1.1307e+07	1.1307e+07	**98.90%**
MSD	2.1068e+05	**-4.2070e+05**	**-4.2077e+05**	97.80%

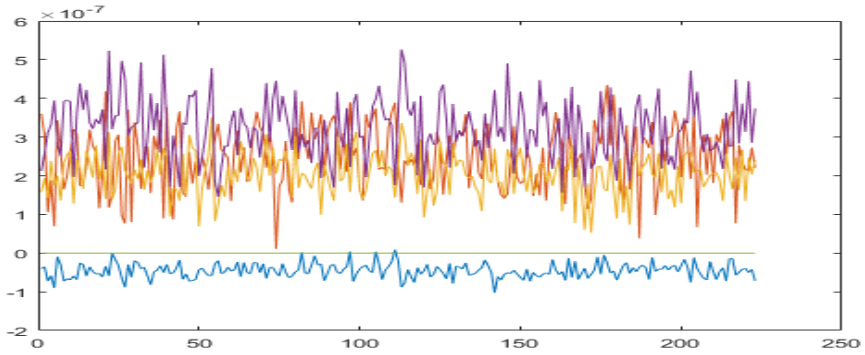
Table 3: Models performance comparison for breast cancer dataset.

Model	Performance metrics			Accuracy
	Log-likelihood	AIC	BIC	
DM	-3.1532e+03	6.3383e+03	6.4111e+03	91.00%
GDM	-3.5300e+03	5.1000e+03	5.1910e+03	93.00%
MBL	-2.4137e+05	4.8414e+05	4.8387e+05	**98.00%**
MSD	-1.7277e+05	**3.4637e+05**	**3.4621e+05**	**98.00%**

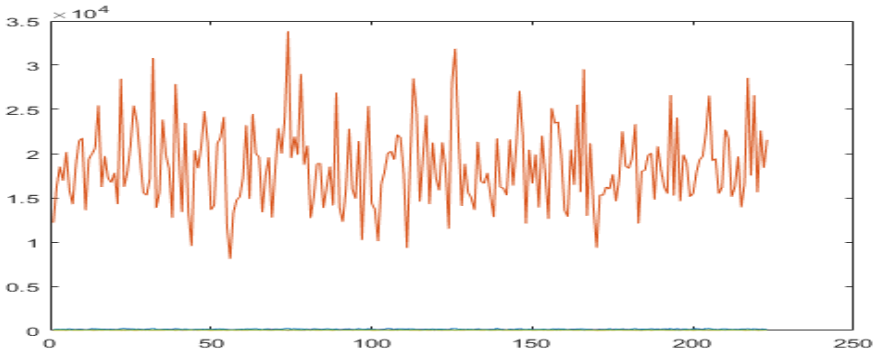
fine needle aspirate (FNA) of a breast mass. The prediction includes the diagnosis of each case to malignant or benign, based on the symmetry, and the fractal dimension. Figure 5 shows sample images from this dataset. After extracting the features, the eight values have been discretized to be used in our models. The eight real-valued features computed for each cell nucleus are; 1-radius (mean of distances from the center to points on the perimeter), 2-texture (standard deviation of gray-scale values), 3-perimeter, 4-area, 5-smoothness (local variation in radius lengths), 6-compactness, 7-concavity (severity of concave portions of the contour), 8-concave points (number of concave portions of the contour).

The prediction results for this dataset are shown in Table 3. As observed that DM and GDM based regression models are not the best fitted models for the prediction of breast cancer shown by the relatively lower accuracy. On the other hand, both MBL and MSD based regression models perform similarly in terms of accuracy. Furthermore, MSD has a lower AIC and BIC, thus, MSD based regression model is better for breast cancer diagnosis dataset.

Figure 6 and Figure 7 illustrate the predicted values for the Breast Cancer dataset, including the three predicted values for each observation using the proposed MBL and MSD based regression models, respectively. As shown in the figures, both proposed models perform well for the prediction of Y values, which is illustrated by having similar behavior for the actual and predicted values for all tested observations.



(a) Y_Train



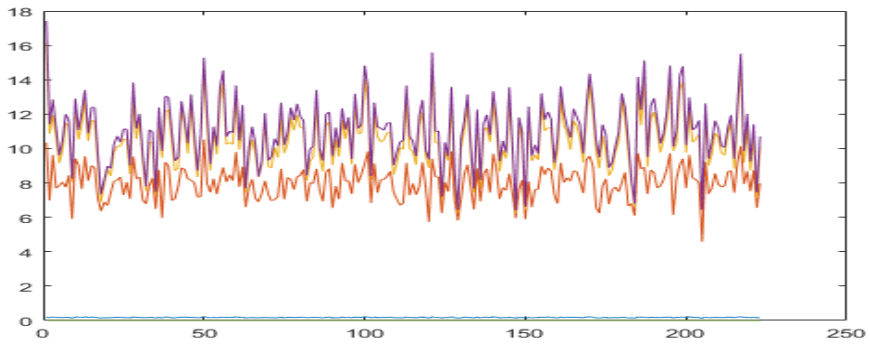
(b) Y-Test

Figure 3: Comparison of test values and the predicted values of Y using MBL-based regression model for Stress Echocardiography dataset.

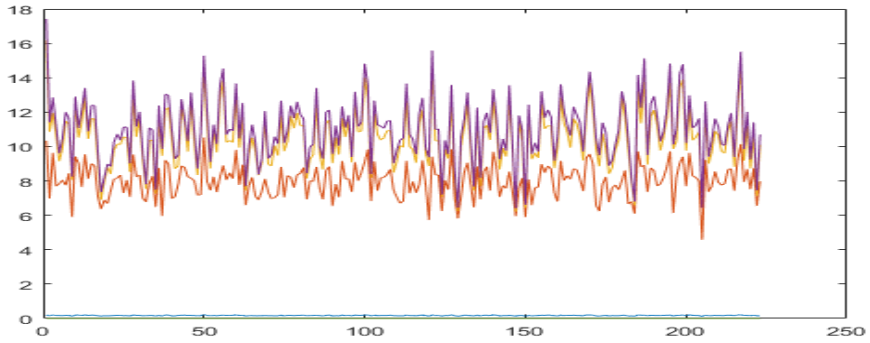
2.2.6 Diagnosis of Diabetes

The Pima Indians Diabetes dataset (DD) [59] dataset used for this application, which is publicly available to download[4]. The objective of this application is to evaluate the efficiency of the proposed models in the problem of diagnosing diabetes. The dataset contains 2,000 observations and nine variables with no missing values reported. The variables in the considered dataset are based on personal data, such as age, the number of pregnancy times, and the results of medical examinations, *e.g.*, blood pressure, body mass index, the result of glucose tolerance test, etc. The analysis aims to predict whether a patient was diabetes positive or not (represented in our experiments by positive count values of 1 and 2, respectively). The dataset consists of a variety of ranges of each feature spanning all individuals. The prediction results

[4]https://www.kaggle.com/uciml/pima-indians-diabetes-database/downloads/pima-indians-diabetes-database.zip/1



(a) Y_Predict



(b) Y_Test

Figure 4: Comparison of test values and the predicted values of Y using MSD-based regression model for Stress Echocardiography dataset.

for this dataset are shown in Table 4. According to the results, DM and GDM-based regression models have a smaller likelihood and relatively lower accuracy. On the other hand, both MSD and MBL outperform the other models, with the MBL regression model has the highest accuracy on this dataset of 99%.

As Figure 8 illustrates, the predicted values using the proposed MBL-based regression model are similar to the actual ones (*i.e.* Y_Test). However, using the MSD-based regression model, the prediction is between 1 and 2. Thus, for predicting if a patient has diabetes or not, the values were rounded to the closest integer. That is, assumed that if the predicted value is greater than 1.5, the diagnosis is negative (actual value is 2); otherwise, the patient has diabetes (actual value is 1).

Table 4: Models performance comparison for Pima Indians Diabetes dataset.

Model	Performance metrics			Accuracy
	Log-likelihood	AIC	BIC	
DM	-622.6466	1.2653e+03	1.3066e+03	92.00%
GDM	-31612.78	6.3664e+04	6.3358e+04	94.50%
MBL	-1.7917e+06	3.5843e+06	3.5842e+06	**99.00%**
MSD	2.8160e+04	**-5.5400e+04**	**-5.5425e+04**	97.75%

Table 5: Comparing the proposed regression models performance to the State-of-the-Art.

DATA SETS	Algorithms	Accuracy
RS	CASI[61]	80.00%
RS	Dirichlet-Multinomial Regression (DM) [2]	94.00%
RS	Generalized Dirichlet-Multinomial Regression (GDM) [2]	95.00%
RS	Proposed model 1 : **MBL-based regression model**	**98.00**%
RS	Proposed model 2 : **MSD-based regression model**	**95.75**%
SEG	CART [57]	95.00%
SEG	Hidden Markov Model (HMM) [62]	93.20%
SEG	Proposed model 1 : **MBL-based regression model**	**98.90**%
SEG	Proposed model 2 : **MSD-based regression model**	**97.80**%
BCD	Logistic Regression[63]	92.10%
BCD	Artificial Neural Networks (ANNs) [64]	96.00%
BCD	Artificial Neural Net Input Gain Measurement Approximation	90.00%
BCD	Proposed model 1: **MBL-based regression model**	**98.00**%
BCD	Proposed model 2: **MSD-based regression model**	**98.00**%
DD	Artificial neural net input gain measurement approximation[65]	71.00%
DD	An Early Neural Network Model(ADAP)[66]	76.00%
DD	Decision Tree [67]	72.00%
DD	ID3 Decision Tree [67]	80.00%
DD	General Regression Neural Network [68]	80.21%
DD	KNN . [68]	77.00 %
DD	Proposed model 1: **MBL-based regression model**	**99.00**%
DD	Proposed model 2: **MSD-based regression model**	**97.75**%

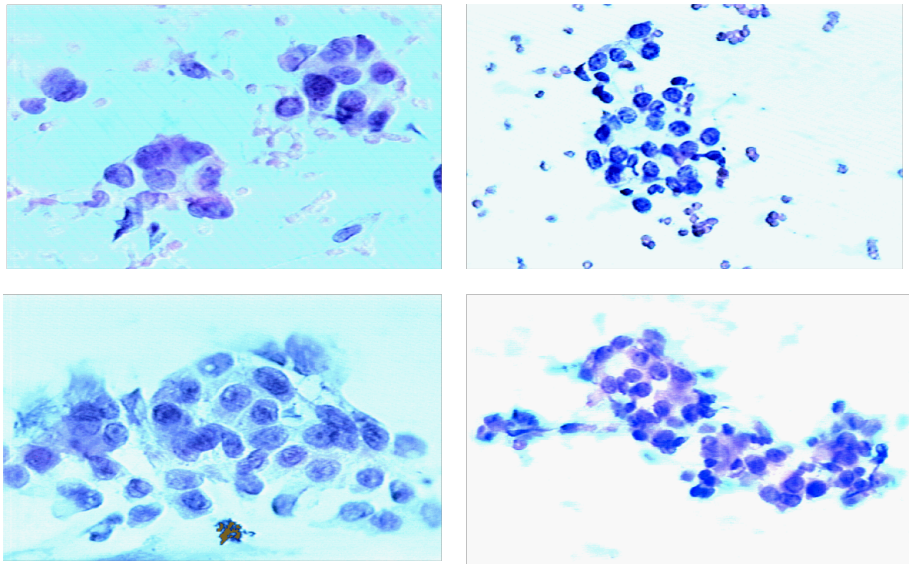
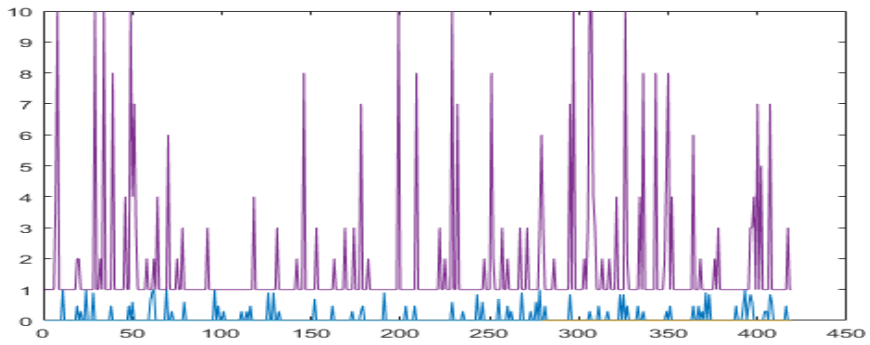


Figure 5: Sample images from the Breast Cancer dataset.

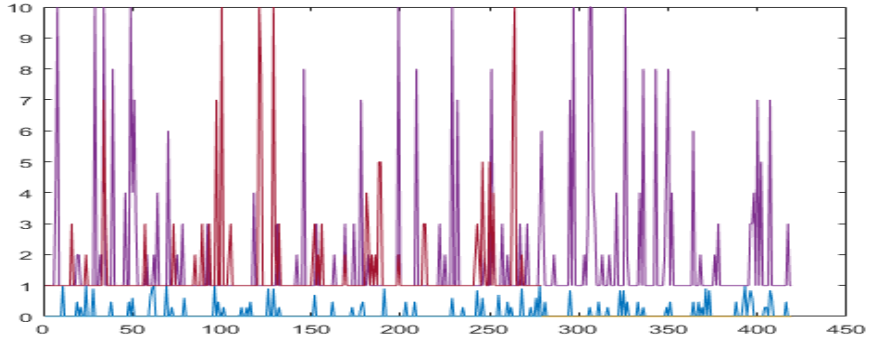
2.2.7 Comparison with Other Methods from the Literature

Recently, a large number of models have been proposed in the literature to perform medical diagnosis efficiently and accurately. In this section, we review the published results for other methods that considered the same datasets used in our experiments. A comparative study between the proposed models and other approaches from the state-of-the-art is depicted in Table 5. From the results of this table it noticed that our proposed approach is competitive to the most successful approaches.

For instance, three different algorithms have been previously implemented to analyze the RNA-seq dataset, including the two with a similar approach (*i.e.*, DM, and GDM-based regression models [2]), and CASI [61]. SEG dataset has been considered using classification and Regression Trees (CART) [57] and Hidden Markov Model (HMM) [62], which are well-known approaches, however, our proposed models achieved the highest accuracy of prediction. Similarly, comparing the results of previous algorithms such as logistic regression [63] and two models based on neural networks [64] implemented on the BCD dataset, our proposed models have the highest accuracy. Furthermore, while the average accuracy of diabetes diagnosis on DD dataset ranges between 71-80%, obtained using previous methods such as logistic regression [63], different neural network models [66, 68], decision trees [67] and KNN [68], our proposed approaches achieve a superior performance of 97.75% and 99% for



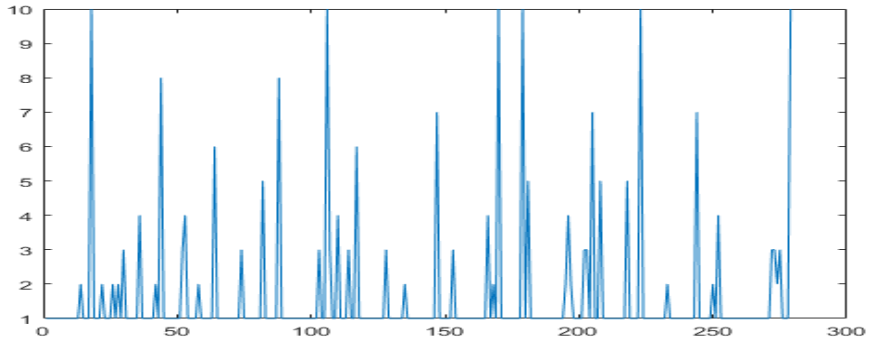
(a) Y_Train



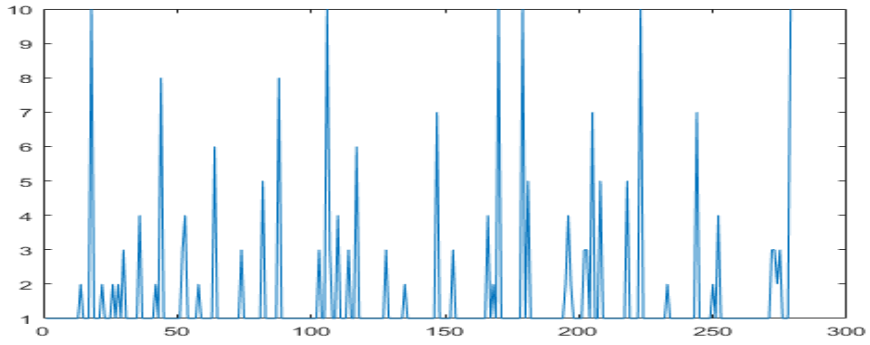
(b) Y-Test

Figure 6: Comparison of test values and the predicted values of Y using MBL-based regression model for Breast Cancer dataset.

the proposed regression models based on MSD and MBL, respectively.

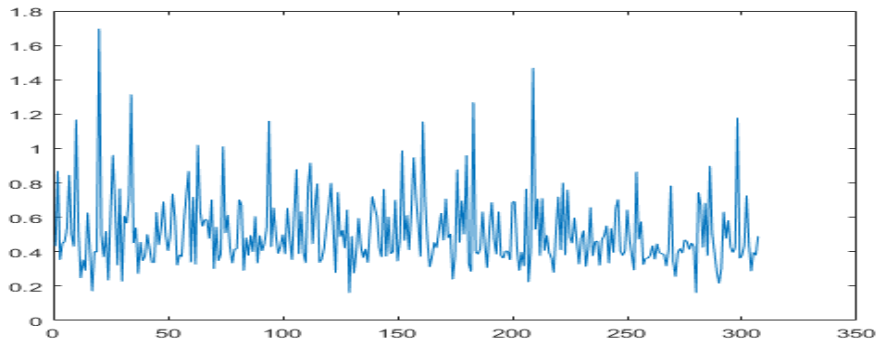


(a) Y_Train

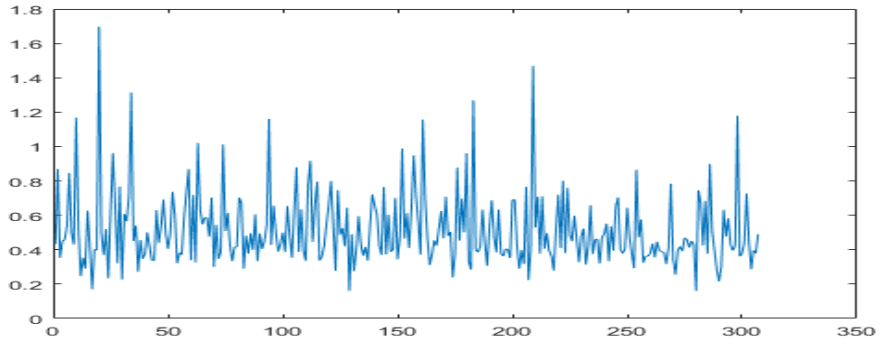


(b) Y-Test

Figure 7: Comparison of test values and the predicted values of Y using MSD-based regression model for Breast Cancer dataset.



(a) Y_Train



(b) Y-Test

Figure 8: Comparison of test values and the predicted values of Y using MSD regression model for Pima Indians Diabetes dataset.

Chapter 3

Distribution-based Regression for Semi-Bounded Data

In this chapter, we focus on modeling and prediction in the case of semi-bounded data which are naturally generated by many real-life applications. Distribution-based regression models using efficient generative models based on Inverted Dirichlet (IDR), Generalized Inverted Dirichlet (GID), and Inverted Beta-Liouville (IBL) distributions has been proposed.

To introduce our model, the response distributions and propose the link functions has been explained. The model parameters are calculated by maximum likelihood approach and to measure the goodness of our model performance, some information measures such as AIC, BIC, and MSE has been used. The efficiency of the proposed models in analyzing real data has shown in the last part.

The structure of the rest of this chapter is as follows. The proposed distribution-based regression models are presented in section 3.1, with all the details about parameters estimation approach and link functions. Section 3.2 presents the experimental results.

3.1 Proposed Regression Models

In this section, the details of the proposed models for IDR regression, GID regression, and IBL regression has explained. For each proposed model, the properties of the fitting distribution was discussed first, then the link functions has been derived and

estimate the parameters using the maximum likelihood approach for each distribution.

3.1.1 The Considered Distributions

Inverted Dirichlet Distribution

The Inverted Dirichlet distribution has been introduced by Tiao and Cuttman [17] for the first time to allow several symmetric and asymmetric modes [69, 70]. If a D-dimensional positive vector $X = (x_1, x_2, ..., x_D)$ follows an inverted Dirichlet distribution, the joint density function is given by [69]:

$$\mathcal{ID}(\mathbf{X}|\theta) = \frac{\Gamma(|\vec{\alpha}|)}{\prod_{d=1}^{D+1}\Gamma(\alpha_d)} \prod_{d=1}^{D} x_d^{\alpha_d - 1}\Big(1 + \sum_{d=1}^{D} x_d\Big)^{-|\vec{\alpha}|} \tag{38}$$

with the condition of $x_d > 0, d = 1, 2, ..., D$, $\vec{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_{D+1})$ and $|\vec{\alpha}| = \sum_{d=1}^{D+1} \alpha_d$ where $\alpha_d > 0$ and $d = 1, 2, ..., D+1$. The mean and variance of the Inverted Dirichlet distribution are given by [69]:

$$E(x_d) = \frac{\alpha_d}{\alpha_{D+1} - 1} \tag{39}$$

$$Var(x_d) = \frac{\alpha_d(\alpha_d + \alpha_{D+1} - 1)}{(\alpha_{D+1})^2(\alpha_{D+1} - 2)} \tag{40}$$

and the covariance between x_d and X_n is :

$$Cov(x_d, X_n) = \frac{\alpha_d \alpha_n}{(\alpha_{D+1} - 1)^2(\alpha_{D+1} - 2)} \tag{41}$$

Inverted Dirichlet distribution provides a good modeling and powerful analytic tool of positive vectors [69]. The inverted Dirichlet choice is inspired by its excellent performance and statistical properties, namely its versatility towards approximating many shapes [69].

Generalized Inverted Dirichlet distribution

The inverted Dirichlet has the downside of having a very restrictive and purely positive covariance structure [17]. Generalized Inverted Dirichlet (GID) has a more general covariance than the Inverted Dirichlet.

If a D-dimensional positive vector $X = (x_1, x_2, ..., x_D)$, with $X > 0$, follows the Generalized Inverted Dirichlet distribution, then:

$$\mathcal{GID}(\mathbf{X}|\theta) = \prod_{d=1}^{D} \frac{\Gamma(\alpha_d + \beta_d)}{\Gamma(\alpha_d)\Gamma(\beta_d)} \frac{x_d^{\alpha_d - 1}}{\left(1 + \sum_{d=1}^{D} x_d\right)^{\gamma_d}} \tag{42}$$

where $\theta = (\alpha_1, \alpha_2, ...\alpha_D, \beta_1, \beta_2, ..., \beta_D)$ where $\gamma_d = \beta_d + \alpha_d - \beta_{d+1}$ and $\beta_{D+1} = 0$ for $d = 1, \ldots, D$.

The mean, variance and co-variance of GID are given by the following formulas:

$$E(x_d) = \frac{\alpha_d}{\beta_d - 1} \tag{43}$$

$$Var(x_d) = \frac{\alpha_d(\alpha_d + \beta_d - 1)}{(\beta_d - 2)(\beta_d - 1)^2} \tag{44}$$

$$Cov(x_d, X_n) = \frac{\alpha_d \alpha_n}{(\beta_d - 2)(\beta_d - 1)^2} \tag{45}$$

Inverted Beta-Liouville distribution

The inverted Dirichlet distribution has a very rigid covariance structure, which limits its versatility considerably. Inverted Beta-Liouville distribution, on the other hand, has shown in recent studies to be very efficient in modeling positive vectors. As a special case, the IBL distribution includes the inverted Dirichlet distribution and can, therefore, provide more flexibility and better fitting of the data [18, 71]. If a D-dimensional vector $X = (x_1, x_2, ..., x_D)$ is drawn from the IBL distribution, then we have:

$$\begin{aligned} \mathcal{IBL}(X|\theta) = {} & \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \prod_{d=1}^{D} \frac{x_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \times \lambda^{\beta} \Big(\sum_{d=1}^{D} x_d \Big)^{\alpha - \sum_{d=1}^{D} \alpha_d} \\ & \times \left(\lambda + \sum_{d=1}^{D} x_d \right)^{-(\alpha + \beta)} \end{aligned} \tag{46}$$

where $\theta = \{\alpha_1, ..., \alpha_d, \alpha, \beta, \lambda\}$, $X > 0$ and $\alpha, \beta, \lambda > 0$. The mean, variance and co-variance of IBL distribution are given as follows [18]:

$$E(x_d) = \frac{\lambda \alpha \alpha_d}{\sum_{d=1}^{D} \alpha_d \beta - 1} \tag{47}$$

$$Var(x_d) = \frac{\lambda^2\alpha(\alpha+1)^2\alpha_d}{(\beta-1)(\beta-2)(\sum_{d=1}^{D}\alpha_d)(\sum_{d=1}^{D}\alpha_d+1)} - \frac{\lambda^2\alpha^2\alpha_d^4}{(\beta-1)^2(\sum_{d=1}^{D}\alpha_d)^4} \tag{48}$$

$$Cov(X_l, X_j) = \frac{\alpha_l\alpha_j}{\sum_{d=1}^{D}\alpha_d}\Big[\frac{\lambda^2\alpha(\alpha+1)}{(\beta-1)(\beta-2)(\sum_{d=1}^{D}\alpha_d)} - \frac{\lambda^2\alpha^2}{(\beta-1)^2(\sum_{d=1}^{D}\alpha_d)}\Big] \tag{49}$$

3.1.2 Link Functions

The link function [44] is the reverse of any distribution-related cumulative distribution function. This function provides the relationship between the linear projection and the mean of the distribution function [45].

Link functions for Inverted Dirichlet distribution

For Inverted Dirichlet distribution-based regression, the relation between the parameters and the p-dimensional co-variate vector $X = (x_1, \ldots, x_p)$, can be written in the following forms:

$$\alpha_i = \lambda_1(\alpha_i x_1 + \alpha_i x_2 + \ldots + \alpha_i x_p), \qquad i = 1, \ldots, d$$
$$\alpha_{D+1} = \lambda_2(\alpha_{D+1} x_1 + \alpha_{D+1} x_2 + \ldots + \alpha_{D+1} x_p)$$

For finding the $\lambda(\mu_i)$ the following procedure has been considered:

$$\lambda(\mu_i) = X_i^T\theta \qquad i = 1, \ldots, d \tag{50}$$

where μ_i is the mean of X_i, and θ is a vector of regression parameters. Thus:

$$logit(\mu_i) = \log\Big(\frac{\mu_i}{1-\mu_i}\Big), \qquad i = 1, \ldots, d \tag{51}$$

and for *logit* link function we have the following :

$$\Pi_j(x) = \frac{\exp(\theta^T X_j)}{1+\sum_{j=1}^{n-1}\exp(\theta^T X_j)} \qquad j = 1, \ldots, p \tag{52}$$

Thus, for the Inverted Dirichlet model, the following equations has been considered:

$$\lambda_1(\mu_i) = X_i^T\alpha_i$$
$$\lambda_2(\mu_i) = X_i^T\alpha_{D+1} \tag{53}$$

Link functions for Generalized Inverted Dirichlet distribution

For GID, to link the parameter $\vartheta = \{\alpha_i, \beta_i\}$ to the p-dimensional covariates vector X, as:

$$\alpha_i = \lambda_1(\alpha_i x_1 + \alpha_i x_2 + \ldots + \alpha_i x_p) \tag{54}$$

$$\beta_i = \lambda_2(\beta_i x_1 + \beta_i x_2 + \ldots + \beta_i x_p), \qquad i = 1, \ldots, d \tag{55}$$

For finding the $\rho(\mu_i)$ the following procedure has been followed:

$$\rho(\mu_i) = X_i^T \vartheta, \qquad i = 1, \ldots, d \tag{56}$$

then we have:

$$\rho_1(\mu_i) = X_i^T \alpha_i \tag{57}$$

$$\rho_2(\mu_i) = X_i^T \beta_i \tag{58}$$

Link function for Inverted Beta-Liouville distribution

For Inverted Beta-Liouville distribution-based regression, the relation between the regression coefficient θ and the p-dimensional co-variate vector $X = (x_1, \ldots, x_p)$, can be written as follows:

$$\begin{aligned}
\alpha_i &= g_1(\alpha_i x_1 + \alpha_i x_2 + \ldots + \alpha_i x_p), \qquad i = 1, \ldots, d \\
\alpha &= g_2(\alpha x_1 + \alpha x_2 + \ldots + \alpha x_p), \\
\beta &= g_3(\beta x_1 + \beta x_2 + \ldots + \beta x_p), \\
\gamma &= g_4(\gamma x_1 + \gamma x_2 + \ldots + \gamma x_p).
\end{aligned} \tag{59}$$

For finding the $g(\mu_i)$ we consider the following procedure:

$$g(\mu_i) = X_i^T \theta \qquad i = 1, \ldots, d \tag{60}$$

Thus, we have the following link functions for Inverted Beta-Liouville distribution:

$$\begin{aligned}
g_1(\mu_i) &= X_i^T \alpha_i \\
g_2(\mu_i) &= X_i^T \alpha \\
g_3(\mu_i) &= X_i^T \beta \\
g_4(\mu_i) &= X_i^T \gamma
\end{aligned} \tag{61}$$

3.1.3 Parameter Estimation

To estimate the regression coefficients parameters, the Maximum Likelihood Estimate (MLE) technique [48] have been used to find the best coefficient for predicting with our models. Maximum likelihood estimation [49, 50] is a popular method that tries to discover the most probable model that provides the observed result. The maximum likelihood parameter estimates for inverted Dirichlet, Generalized Inverted Dirichlet, and Inverted Beta-Liouvell distribution has been obtained by maximizing the log-likelihood. Let $\mathcal{X} = \{X_1, \ldots, X_N\}$ be a dataset with N instances, the estimation of the parameters is based on maximizing the log-likelihood as follows:

$$\Theta^{(t+1)} = \arg\max_{\Theta} \sum_{j=1}^{N} \log(p(X_j|\Theta)) \tag{62}$$

The log-likelihood for IDR is given by:

$$\begin{aligned}\mathcal{L}_n(\mathcal{X}|\theta) = &\sum_{d=1}^{D} ln(\lambda(\mu_i)) + ln(\Gamma(|\vec{\alpha}|)) - \sum_{d=1}^{D+1} ln[\Gamma(\alpha_d)] + \sum_{d=1}^{D} ln(x_d^{\alpha_d - 1}) \\ &- |\bar{\alpha}|\Big[ln(\sum_{d=1}^{D} x_d) + ln(1 + \frac{1}{\sum_{d=1}^{D} x_d})\Big]\end{aligned} \tag{63}$$

The log-likelihood of GID for N independent data points is computed as follows:

$$\begin{aligned}\mathcal{L}_n(\mathcal{X}|\theta) = \sum_{d=1}^{D}\sum_{i=1}^{N} \Bigg[&ln(\lambda(\mu_i)) + ln(\Gamma(\alpha_d + \beta_d) - \big(ln(\Gamma(\alpha_d)) + ln(\Gamma(\beta_d))\big) \\ &+ [ln(x_d^{\alpha_d - 1}) - ln(1 + \sum_{i=1}^{N} X_i)^{\gamma_d}]\Bigg]\end{aligned} \tag{64}$$

and the log-likelihood for IBL follows the formula :

$$\begin{aligned}\mathcal{L}_n(\mathcal{X}|\theta) = \sum_{i=1}^{N} ln(g(\mu_i)) + \Bigg[&\big(ln(\sum_{d=1}^{D} \alpha_d) + ln(\Gamma(\alpha + \beta)) - ln(\Gamma(\alpha)) - ln(\Gamma(\beta))\big) \\ &+ \sum_{d=}^{D}(-ln(\Gamma(\alpha_d)) + ln(x_d^{\alpha_d - 1})) + ln(\gamma^{\beta}) + ln(\sum_{d=1}^{D} x_d)^{\alpha - \sum_{d=1}^{D} \alpha_d} \\ &+ ln\big(\gamma + \sum_{d=1}^{D} x_d^{-(\alpha+\beta)}\big)\Bigg]\end{aligned} \tag{65}$$

Maximizing the log-likelihood function is done by taking the first partial derivatives and solve for the parameters. However, for the three proposed models, closed-form solutions do not exist. Thus, the process requires a Newton-Raphson optimization that iterates between scoring steps based on the present values and an update of the parameters. such that:

$$\Theta^{(t+1)} = \Theta^{(t)} - H_{\Theta}^{-1} G_{\Theta} \tag{66}$$

The first and second order derivatives of the log-likelihood function with respect to θ are shown by G and H where G is the gradient, and H is the Hessian matrix. The complete derivations required to estimate the parameters of IDR, GID, and IBL models has been shown in the following section.

3.1.4 MLE for the proposed models

- **The derivatives to estimate the IDR-based model parameters** The first derivatives of IDR log-likelihood with respect to the regression coefficient are given by:

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_d} = \sum_{i=1}^{n} \lambda_1(\mu_i) \Big[\psi(\vec{\alpha}) - \psi(\alpha_d) + log\Big(\frac{x_d}{1 + \sum_{d=1}^{D} x_d}\Big)\Big] \tag{67}$$

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_{D+1}} = \sum_{i=1}^{n} \lambda_1(\mu_i) \Big[\psi(\vec{\alpha}) - \psi(\alpha_D + 1) + log\Big(\frac{1}{1 + \sum_{d=1}^{D} x_d}\Big)\Big] \tag{68}$$

According to the Newton-Raphson method, the second-order derivatives should calculate

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_{d1} \partial \alpha_{d2}} = \sum_{i=1}^{D} \lambda_2''(\mu_i)[\,\psi'(\vec{\alpha}) - \psi'(\alpha_d)] \tag{69}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \alpha} = \sum_{i=1}^{D} \lambda_1''(\mu_i)[\,\psi'(\vec{\alpha})] \tag{70}$$

- **The derivatives to estimate the GID-based model parameters** The first derivatives of GID log likelihood function with respect to $\alpha_i, i = 1, \ldots, d$ and $\beta_i, i = 1, \ldots, d$ are:

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\vartheta)}{\partial \alpha_i} = \rho_1'(\mu_i)[\psi(\alpha_i + \beta_i) - \psi(\alpha_i) + logX_i - log(1 + X_i)] \tag{71}$$

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\vartheta)}{\partial \beta_i} = \rho_2'(\mu_i)[\psi(\alpha_i + \beta_i) - \psi(\alpha_i) - log(1 + X_i)] \tag{72}$$

The second derivatives of GID are as follows :

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_i \beta_i} = \rho_1''(\mu_i)\rho_2''(\mu_i)[\psi'(\alpha + \beta) - log(1 + X_i)] \tag{73}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \alpha_i} = \rho_1''(\mu_i)[\psi'(\alpha_i + \beta_i) - \psi'(\alpha_i)] \tag{74}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \beta_i} = \rho_1''(\mu_i)[\psi'(\alpha_i + \beta_i)] \tag{75}$$

- **The derivatives to estimate the IBL-based model parameters** The first derivatives of IBL log likelihood function are given by

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_d} = \sum_{i=1}^{n} g_1'(\mu_i)\Big[log(x_d) - log(\sum_{d=1}^{D} x_d) + \psi(\sum_{d=1}^{D} \alpha_d) - \psi(\alpha_d)\Big] \tag{76}$$

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha} = \sum_{i=1}^{n} g_2'(\mu_i)\Big[log(\sum_{d=1}^{D} x_d) + log(\sum_{d=1}^{D} x_d + \gamma) + \psi(\alpha + \beta) - \psi(\alpha)\Big] \tag{77}$$

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \beta} = \sum_{i=1}^{n} g_3'(\mu_i)\Big[log\gamma - log(\sum_{d=1}^{D} x_d + \gamma) + \psi(\alpha + \beta) - \psi(\beta)\Big] \tag{78}$$

$$\frac{\partial \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \gamma} = \sum_{i=1}^{n} g_4'(\mu_i)\Big[\frac{\beta}{\gamma} - \frac{\alpha + \beta}{\gamma + \sum_{d=1}^{D} x_d}\Big] \tag{79}$$

According to the Newton-Raphson method, the second-order derivatives should calculate

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_{d1} \partial \alpha_{d2}} = \begin{cases} g_1''(\mu_i)\big[\psi'(\sum_{d=1}^{D} \alpha_d) - \psi'(\alpha_d)\big], \text{if} \quad d_1 = d_2, \\ \\ g_1''(\mu_i)\psi'(\sum_{d=1}^{D} \alpha_d) \qquad\qquad\qquad\qquad\qquad \text{otherwise}, \end{cases} \tag{80}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \alpha} = g_2''(\mu_i)[\,\psi'(\alpha + \beta) - \psi'(\alpha)] \tag{81}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \beta} = g_3''(\mu_i)[\,\psi'(\alpha+\beta) - \psi'(\beta)] \tag{82}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial^2 \gamma} = g_3''(\mu_i)[-\frac{\beta}{\gamma^2} + \frac{\alpha+\beta}{(\gamma+\sum_{d=1}^{D} x_d)^2}] \tag{83}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha \partial \beta} = g_2''(\mu_i) g_3''(\mu_i)[\,\psi'(\alpha+\beta)] \tag{84}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha \partial \alpha_d} = \frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_d \partial \beta} = \frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha_d \partial \gamma} = 0 \tag{85}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \alpha \partial \gamma} = -[g_2''(\mu_i) g_4''(\mu_i)][\,\frac{1}{\gamma+\sum_{d=1}^{D} x_d}] \tag{86}$$

$$\frac{\partial^2 \mathcal{L}_n(\mathcal{X}|\theta)}{\partial \beta \partial \gamma} = [g_3''(\mu_i) g_4''(\mu_i)][\,\frac{1}{\gamma} - \frac{1}{\gamma+\sum_{d=1}^{D} x_d}] \tag{87}$$

3.1.5 Prediction

Another common task is to predict a numerical target value, given a set of features called predictors. Regression is used to perform this task. For our proposed regression model, the following formula has been used:

$$\hat{Y} = h_\eta X = \eta^T X \tag{88}$$

where for IDR regression with $\theta = (\alpha_1, ..., \alpha_d, \alpha_{D+1})$ regression coefficient. $\eta = \alpha_i \alpha_{D+1}$. For IDR the regression coefficient is equal to $\theta = (\alpha_1, ..., \alpha_d, \beta_1, ..., \beta_d)$. Moreover, assumed that $\eta = \alpha_i \beta_i$, and $\alpha_i \alpha \beta \lambda$ for the GID and IBL regression models, respectively.

The initialization of the parameters was performed using random values. Then, using the maximum likelihood, the parameters are updated in order to obtain their estimates with respect to a given dataset. The complete learning algorithm is summarized in (Algorithm 2).

Algorithm 2 The complete learning algorithm

1. **Input** DATA SET $\mathcal{X} = \{W_1, \ldots, W_n\}$ with n independent data points $W_j = (X_i, Y_i)$, where X_i is the count response vector and Y_i is covariate vector.
   Output The final parameters Θ, log-likelihood, predicted Y
2. Split the data by ratio 60:40 for training and testing
3. Initialize the parameters for each model $\Theta^{(0)}$
4. **repeat**
5. Update the parameters $\Theta^{(t)}$ using Eq.(66)
6. Update the link functions
7. calculate the complete log-likelihood using Eq.(63) for IDR or Eq.(64) for GID or Eq.(65) for IBL
8. **until** *convergence*
9. Predict the covariate values of Y using Eq. (88).

3.2 Experimental Results

Our aim in this section is to apply the proposed regression models on real datasets. The performance of the three proposed regression models based on inverted Dirichlet, generalized inverted Dirichlet and inverted Beta-Liouville has been compared. Moreover, the effectiveness of the proposed models compared to two widely used regression models, namely, linear and logistic regressions has been shown.

3.2.1 Data and Performance Measures

The evaluation of each model is based on the Akaike information criterion (AIC) [52], Bayesian information criterion (BIC) [53], and MSE where smaller values for AIC, BIC and MSE indicate that the model has a better performance. Furthermore, we considered the prediction accuracy.

3.2.2 Software Defects Prediction

Software quality control and identification of a flaw or defect in a computer program have become one of the research subjects that has received a lot of attention. Any device failure can result in high costs [72]. It is challenging and difficult to evaluate the quality of complex software systems. Therefore, forecasting program failures is a desirable tool to improve reliability [73, 74]. There are some software complexity evaluation [75] measures such as code size, cyclomatic McCabes, and complexity of Halsteads that could be used for prediction.

Our analysis is carried out on two datasets from the archive of PROMISE data obtained from NASA software projects and its public MDP (Modular Data Processing Toolkit), which are currently used as benchmark datasets in this research area. The metrics or features of each dataset are five different lines of calculation of code, three metrics of McCabe, four measures of base Halstead, eight measures of derived Halstead, and a branch-count. A binary variable classifies the datasets to show whether or not the module is faulty. PC1 is a NASA spacecraft instrument software which considers functions flight software for earth-orbiting satellite. JM1 is a predictive ground-based system in real-time. Both softwares are written in "C". Table 6 presents the results of the three tested models, where compared them based on AIC, BIC, MSE, and accuracy. As it shown, the IBL based model has the smallest AIC, BIC, MSE and highest accuracy of 98 % and Table 7 shows the results of applying regression models on the JM1 dataset and show that IBL regression is the best model for this dataset in terms of accuracy. The IBL based model outperforms all the tested models with an accuracy of 98% compared to 67-82% for the other models. As Tables 6 and 7 shown our proposed model based on IBL outperforms all the tested models. Moreover, the proposed regression model based on GID performs better than the linear regression.

3.2.3 Age prediction

The second real-world application that considered is age prediction from facial images using the UTKFace [76] data sets to evaluate our three proposed models. Recently, research on face and age prediction has become a trendy topic in machine learning [77]. Approaches are falling mostly into two groups, physical models and prototype-based approaches. Examples include predicting ageing biological process and body

Table 6: Models performance comparison for software defects prediction in PC1 dataset.

Model	Performance metrics			Accuracy
	MSE	AIC	BIC	
ID	1.2084	-264744.86	-264739.64	67.00%
GID	1.09	-446644.53	-446749.74	82.00%
IBL	5.83026e+29	- 567644.53	-556849.98	**98.00%**
Linear regression	1.23	-356765.12	-356899.23	78.11%
logistic regression	0.055	-495784.98	-487695.67	86.00 %

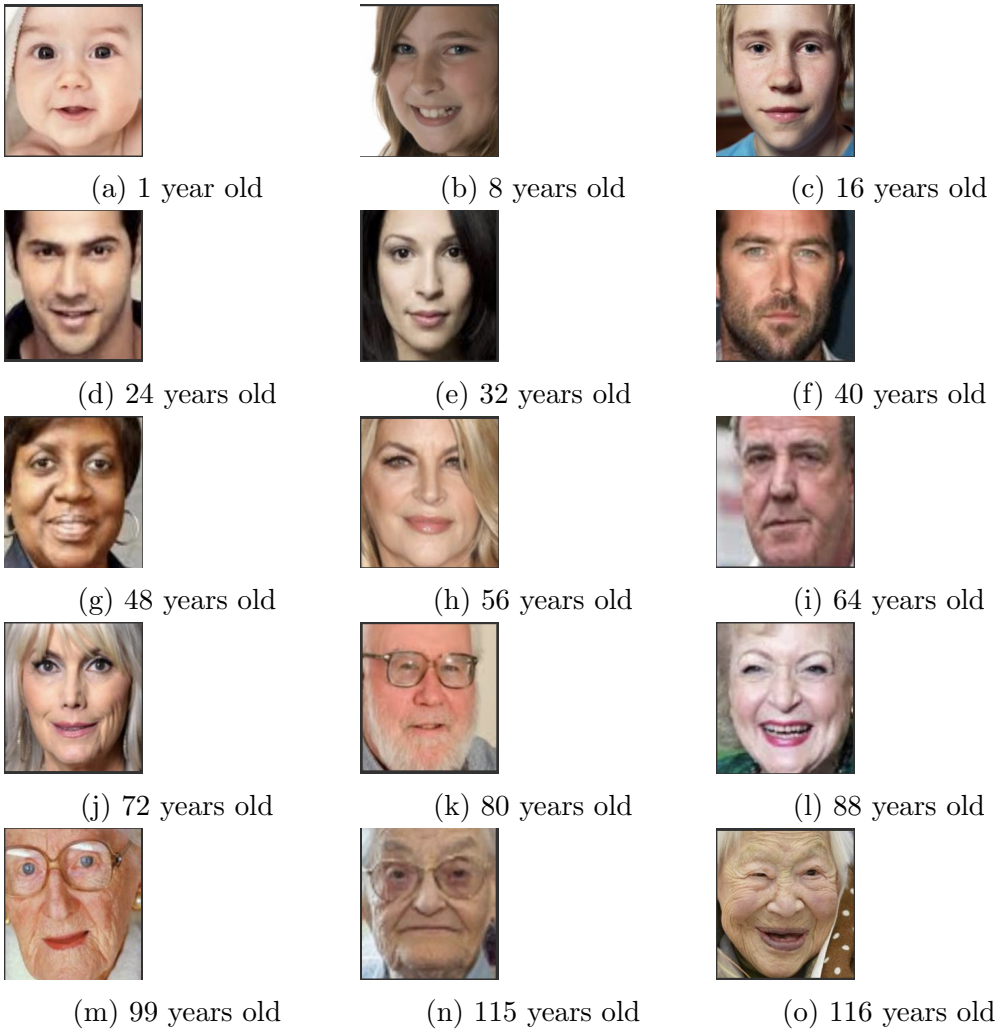
Table 7: Models performance comparison for for software defects prediction in JM1 dataset.

Model	Performance metrics			Accuracy
	MSE	AIC	BIC	
ID	1.58	-155737.96	-155737.76	67.00%
GID	1.25	- 3625146.01	-3625299.215	82.00%
IBL	1.07	-4562146.98	-4562641.35	**98.00%**
Linear regression	1.14	-312115.23	-311399.11	81.24%
logistic regression	1.1	-395784.98	-387695.67	90.00 %

processes such as wrinkles [78], facial structure [79] and muscles [80].

UTKFace dataset is a large-scale face dataset with a wide age range between 0 and 116 years. The dataset is made up of approximately 20,000 face photos with age, ethnicity, and gender annotations (see sample images in Figure 9). The images cover significant variations in pose, expression, brightness, occlusion, frame rate. the 23,675 aligned and cropped face images collected from the UTK dataset; each has a size of 200×200 pixels. For feature extraction and description, the Histogram of oriented gradients (HOG) technique used with a cell size of 64 ended up with a 144-dimensional feature vector for each image.

The prediction results for this dataset are shown in Table 8 demonstrated by the three performance metrics and the overall accuracy. According to the results, with a comparison of the three proposed models, the IBL has the smallest MSE, AIC, and BIC and the highest accuracy. Thus, IBL based regression is the best model for age prediction. Compared to the common linear regression, the proposed IBL regression model has a significantly better performance.



(a) 1 year old (b) 8 years old (c) 16 years old

(d) 24 years old (e) 32 years old (f) 40 years old

(g) 48 years old (h) 56 years old (i) 64 years old

(j) 72 years old (k) 80 years old (l) 88 years old

(m) 99 years old (n) 115 years old (o) 116 years old

Figure 9: Sample images from the UTK dataset

3.2.4 Spam Filtering

Our experiment was conducted on a complex spam dataset developed by Hewlett-Packard Labs from the UCI machine learning repository [81]. Spam filtering[82] is one of the major research fields in the security of information systems. There are serious threats caused by spams or unsolicited bulk communications. As reported in the literature, up to 75–80% of e-mail messages in 2005 and 2009 are spam, which resulted in huge financial losses between $50 and $130 billion [83, 84]. The considered dataset contains 4,601 instances and 58 attributes (57 attributes of continuous input and 1 target mark attribute of nominal class). Around 39.4% (1813 instances) of the e-mails are spam, and 60.6% (2788 e-mails) are not. The attributes are extracted

Table 8: Models performance comparison for Age prediction in UTK dataset.

Model	Performance metrics			Accuracy
	MSE	AIC	BIC	
ID	40.95	-14176410.34	-14175247.95	68.91%
GID	21.04	-89379377.99	-89378215.59	90.40%
IBL	17.58	-98398512.86	-98318289.31	**95.00%**
Linear regression	47.5	-48735647.78	-48756247.78	59.72%

using one of the main methods of information representation in natural language processing called Bag of Words (BoW) [85]. Every e-mail is identified by its words regardless of grammar in this process. Most attributes in the spam base dataset indicate whether the e-mail often contained a particular word or character, 48 features include the percentage of words in the e-mail that match the word. The remaining characteristics are the average length of uninterrupted capital letter series, the length of the most extended uninterrupted capital letter sequence, and the total number of capital letters in the e-mail. The rating of the dataset shows whether or not the e-mail was deemed spam. In our experiments, the dataset was first reduced to 3626 instances to have a balanced case. The prediction results for this dataset are shown in Table 9. IDR and GID based regression models have the same MSE for the prediction of Spam emails. Furthermore, IBL has a lower MSE, AIC, BIC, and highest accuracy. Thus, the IBL based regression model is better for Spam prediction. Compared to the common regression models, the three proposed regression models have significantly higher accuracy.

Table 9: Models performance comparison for Spam filtering.

Model	Performance metrics			Accuracy
	MSE	AIC	BIC	
ID	2.18	-16041.90	-11675.17	94.00%
GID	2.18	-4308416.48	-4308783.21	94.00%
IBL	1.5635e+25	-7348436.84	-7348436.12	**97.00%**
Linear regression	2.32	-2301.56	-2681.32	52.90%
logistic regression	2.10	-4332.81	-4236.12	68.5 %

3.2.5 Disease Diagnosis

Hepatitis diagnosis

Hepatitis, an inflammation of the liver, is commonly caused by viruses [86], but its origin could be other factors, including allergies, autoimmune diseases, or toxic substances. Blood testing is the primary method for diagnosing it. Automatic diagnostic techniques can assist doctors in diagnosing diseases accurately. In this experiment, the proposed regression model applied on a dataset [87], which includes 155 instances and 19 attributes. Features include age, presence of steroid, antivirals administered, fatigue, malaise, anorexia, large liver, firm liver, spleen palpability, presence of spiders, presence of ascites, presence of varices, bilirubin level, alkaline phosphate level, SGOT level, albumin level, protein level, and histology result. We use our models to predict if the patient is alive or not. The prediction results for the considered dataset using the three tested regression models are given in Table 10 reported using the above-mentioned performance metrics. According to the results, one may notice that ID and IBL have approximately similar performance according to accuracy. On the other hand, the proposed IBL has the smallest MSE, AIC, and BIC. Thus, IBL based regression is the best model for predicting Hepatitis. Compared to the common regression models, the proposed GID regression model has slightly better performance, such that its accuracy is 3% higher, where the other two proposed regression models have significantly higher accuracies of 97% and 98% for the ID and IBL based regression models, respectively.

Table 10: Models performance comparison for Hepatitis diagnosis dataset.

Model	Performance metrics			Accuracy
	MSE	AIC	BIC	
ID	1.58	-787270.38	-787270.14	97.00%
GID	3.56	-60877.60	-60929.34	79.00%
IBL	1.03	-85268.60	-85368.78	**98.00%**
Linear regression	2.28	-50157.69	-51469.18	74.11%
logistic regression	2.53	-54165.19	-54149.59	76.7 %

Liver disorder diagnosis

The last dataset used in our experiments is the liver-disorders [88]. This dataset consists of several attributes; the first 5 variables are all blood tests that are considered to be responsive to liver abnormalities that may result from excessive consumption of alcohol. Each observation in the dataset is a single male individual's record. The features are mcv mean corpuscular volume, alkphos alkaline phosphotase,sgpt alanine aminotransferase, sgot aspartate aminotransferase, gammagt gamma - glutamyl transpeptidase, drinks number of half-pint equivalents of alcoholic beverages drunk per day and selector that shows the patient has the disease or not. In the past, the last label (selector) was frequently misinterpreted as a dependent variable describing a liver condition presence or absence. BUPA researchers created the seventh field.

Table 11 shows the prediction results of this dataset. ID and GID based regression models are as good as IBL fitted models for the prediction of liver disorder shown by the relatively lower accuracy. On the other hand, both IDR and GID based regression models perform approximately similarly in terms of accuracy, AIC, BIC, and MSE. Furthermore, the IBL based regression model has the highest accuracy and lowest MSE, AIC, and BIC; thus, it is the best model for diagnosing liver disorder. Table 11 shows that the two standard regression models have lower accuracy, and our proposed models have the highest accuracy and the best prediction results.

Table 11: Models performance comparison for Liver disorder dataset.

Model	Performance metrics			Accuracy
	MSE	AIC	BIC	
ID	4.64	-38952.35	-38790.87	96.00%
GID	4.35	-37480.223	-37499.44	95.00%
IBL	4.90e+53	-55289.45	-55398.22	**98.00%**
Linear regression	5.61	-12457.49	-11469.27	77.5%
logistic regression	5.07	-28765.38	-28549.18	87.0 %

Chapter 4

Conclusion

In this thesis, different regression techniques for count and semi-bounded data has been explored in details. We started our work by introducing two novel regression models for count data based on multinomial Beta-Liouville and multinomial scaled Dirichlet distributions. The two proposed models are mainly motivated by the fact that these distributions offer high flexibility, better fitting, and considerable potential to accurately describe count data compared to previously used models. To validate the performance of these models, the application of assessing the connections and patterns analysis in medical data has been considered. The evaluation is performed by considering different measures that are usually used to evaluate regression models, including model selection criteria such as AIC and BIC, as well as the prediction accuracy. According to the obtained results, our models achieved superior performance supported by higher accuracy of predicting diseases. It could be claimed that these new distribution-based regression models yield better results than the other comparable state-of-the-art methods. Further, three novel regression models for semi-bounded data based on flexible distributions for positive vectors has been proposed, namely, inverted Dirichlet, generalized inverted Dirichlet, and inverted Beta-Liouville. This work has shown that these distributions offer high versatility, better fit, and considerable potential to represent positive vectors accurately compared to linear and logistic regressions. Several real-world applications, including analysis of medical data, spam filtering, age prediction and software defect prediction to validate the efficiency of the proposed models has been considered. The results have demonstrated that our models outperform similar approaches.

Future approaches for research will concentrate on models modifications and improvements to achieve greater precision in regression. Future works could be devoted to the extension of the proposed models to other applications and especially those dealing with time series data. We will focus mainly on regression mixture models of distributions.

Bibliography

[1] Yiwen Zhang, Hua Zhou, Jin Zhou, and Wei Sun. Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13, 2017.

[2] Yiwen Zhang, Hua Zhou, Jin Zhou, and Wei Sun. Regression models for multivariate count data. *Journal of Computational and Graphical Statistics*, 26(1):1–13, 2017.

[3] Aristidis K Nikoloulopoulos and Dimitris Karlis. Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation*, 39(1):172–187, 2009.

[4] Kristian Karstoft, Cecilie Fau Brinkløv, Ida Kær Thorsen, Jens Steen Nielsen, and Mathias Ried-Larsen. Resting metabolic rate does not change in response to different types of training in subjects with type 2 diabetes. *Frontiers in endocrinology*, 8:132, 2017.

[5] Rashedur M Rahman and Farhana Afroz. Comparison of various classification techniques using different data mining tools for diabetes diagnosis. *Journal of Software Engineering and Applications*, 6(03):85, 2013.

[6] Daniel Powers and Yu Xie. *Statistical methods for categorical data analysis*. Emerald Group Publishing, 2008.

[7] Ralf Herbrich, Thore Graepel, Klaus Obermayer, et al. *Regression models for ordinal data: A machine learning approach*. Citeseer, 1999.

[8] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[9] Nizar Bouguila. A model-based approach for discrete data clustering and feature weighting using MAP and stochastic complexity. *IEEE Trans. Knowl. Data Eng.*, 21(12):1649–1664, 2009.

[10] Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

[11] Ricardo Maronna. Alan julian izenman (2008): modern multivariate statistical techniques: regression, classification and manifold learning. *Statistical Papers*, 52(3):733–734, 2011.

[12] David F Andrews. A robust method for multiple linear regression. *Technometrics*, 16(4):523–531, 1974.

[13] Cristian L Bayes, Jorge L Bazán, Catalina García, et al. A new robust regression model for proportions. *Bayesian Analysis*, 7(4):841–866, 2012.

[14] Divya Ankam and Nizar Bouguila. Generalized dirichlet regression and other compositional models with application to market-share data mining of information technology companies. In *Proceedings of the 21st International Conference on Enterprise Information Systems, ICEIS 2019, Heraklion, Crete, Greece, May 3-5, 2019, Volume 1.*, pages 158–166, 2019.

[15] Nizar Bouguila. A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity. *IEEE Transactions on Knowledge and Data Engineering*, 21(12):1649–1664, 2009.

[16] Mohamed Al Mashrgy, Taoufik Bdiri, and Nizar Bouguila. Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted dirichlet mixture models. *Knowledge-Based Systems*, 59:182–195, 2014.

[17] George G Tiao and Irwin Cuttman. The inverted dirichlet distribution with applications. *Journal of the American Statistical Association*, 60(311):793–805, 1965.

[18] Can Hu, Wentao Fan, Ji-Xiang Du, and Nizar Bouguila. A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models. *Neurocomputing*, 333:110–123, 2019.

[19] James E Mosimann. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962.

[20] Nizar Bouguila and Djemel Ziou. On fitting finite dirichlet mixture using ecm and mml. In *International conference on pattern recognition and image analysis*, pages 172–182. Springer, 2005.

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

[22] Nizar Bouguila and Djemel Ziou. Mml-based approach for finite dirichlet mixture estimation and selection. In *International workshop on machine learning and data mining in pattern recognition*, pages 42–51. Springer, 2005.

[23] Tao Wang and Hongyu Zhao. A dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*, 73(3):792–801, 2017.

[24] Nizar Bouguila, Djemel Ziou, and Jean Vaillancourt. Novel mixtures based on the dirichlet distribution: application to data and image classification. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 172–181. Springer, 2003.

[25] Rasmus E Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, pages 545–552. ACM, 2005.

[26] Paulo Guimaraes, Richard Lindrooth, et al. Dirichlet-multinomial regression. *Economics Working Paper Archive at WUSTL, Econometrics*, (0509001), 2005.

[27] Nizar Bouguila. Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):462–474, 2008.

[28] Tzu-Tsung Wong. Alternative prior assumptions for improving the performance of naïve bayesian classifiers. *Data Mining and Knowledge Discovery*, 18(2):183–213, 2009.

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

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

[31] Nuha Zamzami and Nizar Bouguila. Consumption behavior prediction using hierarchical bayesian frameworks. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, pages 31–34. IEEE, 2018.

[32] Nizar Bouguila. Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22(2):186–198, 2010.

[33] Nuha Zamzami and Nizar Bouguila. Text modeling using multinomial scaled dirichlet distributions. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 69–80. Springer, 2018.

[34] Nuha Zamzami and Nizar Bouguila. A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognition*, 2019.

[35] Pantea Koochemeshkian, Nuha Zamzami, and Nizar Bouguila. Flexible distribution-based regression models for count data: application to medical diagnosis. *Journal of Cybernetics and Systems*, 2020, accepted.

[36] Pantea Koochemeshkian, Nuha Zamzami, and Nizar Bouguila. Distribution-based regression for semi-bounded data. In *IEEE International Conference on Systems, Man, and Cybernetics.* submitted, 2020.

[37] Nizar Bouguila. On the smoothing of multinomial estimates using liouville mixture models and applications. *Pattern Anal. Appl.*, 16(3):349–363, 2013.

[38] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

[39] Gianna Serafina Monti, Gloria Mateu-Figueras, and Vera Pawlowsky-Glahn. Notes on the scaled dirichlet distribution. *Compositional Data Analysis*, pages 128–138, 2011.

[40] Nuha Zamzami and Nizar Bouguila. An accurate evaluation of msd log-likelihood and its application in human action recognition. In *7th IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019.

[41] Robin KS Hankin et al. A generalization of the dirichlet distribution. *Journal of Statistical Software*, 33(11):1–18, 2010.

[42] Bromensele Samuel Oboh and Nizar Bouguila. Unsupervised learning of finite mixtures using scaled dirichlet distribution and its application to software modules categorization. In *2017 IEEE International Conference on Industrial Technology (ICIT)*, pages 1085–1090. IEEE, 2017.

[43] Nuha Zamzami, Rua Alsuroji, Oboh Eromonsele, and Nizar Bouguila. Proportional data modeling via selection and estimation of a finite mixture of scaled dirichlet distributions. *Computational Intelligence*, 2019, accepted.

[44] Claire Elayne Bangerter Owen. Parameter estimation for the beta distribution. 2008.

[45] Bani K Mallick and Alan E Gelfand. Generalized linear models with unknown link functions. *Biometrika*, 81(2):237–245, 1994.

[46] CJF Ter Braak. Partial canonical correspondence analysis. In *Classification and related methods of data analysis: proceedings of the first conference of the International Federation of Classification Societies (IFCS), Technical University of Aachen, FRG, 29 June-1 July 1987*, pages 551–558. North-Holland, 1988.

[47] Falk Howar, Bernhard Steffen, Bengt Jonsson, and Sofia Cassel. Inferring canonical register automata. In *International Workshop on Verification, Model Checking, and Abstract Interpretation*, pages 251–266. Springer, 2012.

[48] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[49] Klaus LP Vasconcellos and Francisco Cribari-Neto. Improved maximum likelihood estimation in a new class of beta regression models. *Brazilian Journal of Probability and Statistics*, pages 13–31, 2005.

[50] Philip Paolino. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 9(4):325–346, 2001.

[51] José M Taboada, Javier Rivero, Fernando Obelleiro, Marta G Araújo, and Luis Landesa. Method-of-moments formulation for the analysis of plasmonic nano-optical antennas. *JOSA A*, 28(7):1341–1348, 2011.

[52] Kenneth P Burnham, David R Anderson, and Kathryn P Huyvaert. Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral ecology and sociobiology*, 65(1):23–35, 2011.

[53] Kenneth P Burnham and David R Anderson. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[54] Kenneth P Burnham and David R Anderson. Kullback-leibler information as a basis for strong inference in ecological studies. *Wildlife research*, 28(2):111–119, 2001.

[55] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[56] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[57] Janine Krivokapich, John S Child, Donald O Walter, and Alan Garfinkel. Prognostic value of dobutamine stress echocardiography in predicting cardiac events in patients with known or suspected coronary artery disease. *Journal of the American College of Cardiology*, 33(3):708–716, 1999.

[58] William H Wolberg, W Nick Street, and Olvi L Mangasarian. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. *Analytical and Quantitative cytology and histology*, 17(2):77–87, 1995.

[59] Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

[60] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773, 2010.

[61] Hugues Richard, Marcel H Schulz, Marc Sultan, Asja Nurnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, et al. Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic acids research*, 38(10):e112–e112, 2010.

[62] Kiryl Chykeyuk, David A Clifton, and J Alison Noble. Feature extraction and wall motion classification of 2d stress echocardiography with relevance vector machines. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 677–680. IEEE, 2011.

[63] Andrew I Schein and Lyle Ungar. A-optimality for active learning of logistic regression classifiers. Technical report, 2004.

[64] Hussein A Abbass. An evolutionary artificial neural networks approach for breast cancer diagnosis. *Artificial intelligence in Medicine*, 25(3):265–281, 2002.

[65] Chun-nan Hsu, Dietrich Schuschel, and Ya-ting Yang. The annigma-wrapper approach to neural nets feature selection for knowledge discovery and data mining. *Institute of Information Science*, 1999.

[66] Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

[67] Jianchao Han, Juan C Rodriguez, and Mohsen Beheshti. Diabetes data analysis and prediction model discovery using rapidminer. In *2008 Second international*

conference on future generation communication and networking, volume 3, pages 96–99. IEEE, 2008.

[68] Kamer Kayaer and Tulay Yıldırım. Medical diagnosis on pima indian diabetes using general regression neural networks. In *Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP)*, volume 181, page 184, 2003.

[69] Taoufik Bdiri and Nizar Bouguila. Positive vectors clustering using inverted dirichlet finite mixture models. *Expert Systems with Applications*, 39(2):1869–1882, 2012.

[70] Taoufik Bdiri and Nizar Bouguila. Bayesian learning of inverted dirichlet mixtures for SVM kernels generation. *Neural Computing and Applications*, 23(5):1443–1458, 2013.

[71] Kai Wang Fang. *Symmetric multivariate and related distributions*. Chapman and Hall/CRC, 2018.

[72] Naoki Kawashima and Osamu Mizuno. Predicting fault-prone modules by word occurrence in identifiers. In *Software Engineering Research, Management and Applications*, pages 87–98. Springer, 2015.

[73] Alexandre Boucher and Mourad Badri. Predicting fault-prone classes in object-oriented software: an adaptation of an unsupervised hybrid som algorithm. In *2017 IEEE International Conference on Software Quality, Reliability and Security (QRS)*, pages 306–317. IEEE, 2017.

[74] Michael R Lyu et al. *Handbook of software reliability engineering*, volume 222. IEEE computer society press CA, 1996.

[75] Saiqa Aleem, Luiz Fernando Capretz, and Faheem Ahmed. Benchmarking machine learning technologies for software defect detection. *arXiv preprint arXiv:1506.07563*, 2015.

[76] Utkface. available https://susanqq.github.io/UTKFace/.

[77] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

[78] Narayanan Ramanathan and Rama Chellappa. Modeling shape and textural variations in aging faces. In *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8. IEEE, 2008.

[79] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on pattern Analysis and machine Intelligence*, 24(4):442–455, 2002.

[80] Jinli Suo, Xilin Chen, Shiguang Shan, Wen Gao, and Qionghai Dai. A concatenational graph evolution aging model. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2083–2096, 2012.

[81] Spambase UCI Repository data set 1999. *IEEE Transactions on software Engineering*, 1999.

[82] Nizar Bouguila and Ola Amayri. A discrete mixture-based kernel for svms: Application to spam and image categorization. *Inf. Process. Manag.*, 45(6):631–642, 2009.

[83] Yuanchun Zhu and Ying Tan. A local-concentration-based feature extraction approach for spam filtering. *IEEE Transactions on Information Forensics and Security*, 6(2):486–497, 2010.

[84] Levent Özgür and Tunga Güngör. Optimization of dependency and pruning usage in text classification. *Pattern analysis and applications*, 15(1):45–58, 2012.

[85] Claudia Aparecida Martins, Maria Carolina Monard, and Edson Takashi Matsubara. Reducing the dimensionality of bag-of-words text representation used by learning algorithms. In *Proc of 3rd IASTED International Conference on Artificial Intelligence and Applications*, pages 228–233, 2003.

[86] Hepatitis. available at https://www.who.int/features/qa/76/en/e.

[87] Hepatitis. available at https://archive.ics.uci.edu/.

[88] Hepatitis. available at https://www.openml.org/d/8.