Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DATA SEGMENTATION USING MIXTURE REGRESSION MODELS WITH GENERALIZED GAUSSIAN DISTRIBUTION AND K-MEANS

S A V S SAMBHA MURTHY S [1], K. SRINIVASA RAO [2], KUNJAM NAGESWAR RAO [3]

[1]Research Scholar, Department of Computer Science &Systems Engineering, College of Engineering (A), Andhra University, Visakhapatnam, India.
[2]Senior Professor, Andhra University, Visakhapatnam, India.
[3]Professor, Department of Computer Science &Systems Engineering, College of Engineering (A), Andhra University, Visakhapatnam, India.

e-mail : [1]sivakotimurthy45@gmail.com, [2] ksraoau@yahoo.co.in, [3]kunjamnag@gmail.com

ABSTRACT

Data segmentation using mixture regression models gained lot of momentum due to their ready applicability in market analytics, business analytics, financial analytics, supply chain analytics , Human Resource analytics etc. In regression analysis it is customary to assume that error term follows a Gaussian distribution. Gaussian distribution has several drawbacks such as being mesokurtic and the model may not serve well for all types of data. Hence, in this paper we develop data segmentation method using mixture of regression models with Generalized Gaussian Distributed (GGD) errors. The GGD includes leptokurtic, platykurtic and Gaussian distribution as particular cases. The model parameters are estimated using Expectation Maximization (EM) algorithm. The initialization of the parameters is done by using k-means algorithm. The data segmentation algorithm is derived using component maximum likelihood under Bayesian framework. The utility of the proposed algorithm is demonstrated with market segmentation. The performance of the algorithm is evaluated by computing segmentation performance metrics such as accuracy, misclassification rate, precision. It is observed that this method performs much better than the earlier data segmentation methods having Gaussian distributed errors for the data sets having leptokurtic and platykurtic response variables.

Keywords: *Segmentation Methods, Regression Analysis, Generalized Gaussian distribution, Market Analytics*, *Expectation and Maximization Algorithm.*

1. INTRODUCTION

Clusterwise Linear Regression (CLR) is a technique based on the combination of clustering and linear regression. In the literature it is also referred to as regression clustering, switching regression. The objective of CLR is to find a given number of linear functions each approximating a subset of the whole data set by minimizing the overall sum of regression errors. CLR can be considered as extension of linear regression. One linear function is used to fit the whole data set in the linear regression where as CLR approximates the data using more than one linear functions. CLR has been applied to several application domains including customer benefit segmentation [9], market segmentation [10], modeling of the metal inert gas welding process [11], pavement management systems [12], rain fall prediction [13] and PM10 prediction [14].

Wayne S. Desarbo et.al [1] presented a conditional mixture, maximum likelihood methodology for performing clusterwise linear regression. This methodology estimates separate regression functions and membership in K clusters or groups simultaneously. Qiang Long et.al [2] described various methods to solve clusterwise linear regression problems. Ye Chow Kuang et.al [15] presented the first examination of CLR algorithms developed over the past two decades through randomized large-sample testing. Yifan Zhang et.al [16] proposed a new generalized ordinal Bayesian finite mixture regression model for market segmentation which allows simultaneous variable selection within each derived segment and recovers segment profiling using concomintant variables. Ting Li et. Al [17] extended the classical

clusterwise linear regression to incorporate multiple functional predictors by representing the functional coefficients in terms of a functional principal component basis. Kaisa Joki et. Al [18] studied a model and solved the CLR problem by using support vector machines for regression to approximate each cluster. Paul W. Murray et. Al [19] applied data mining methods to identify behavior patterns in historical noisy delivery data in market segmentation. Cathy W.S. Chen et. Al [20] studied a Bayesian approach to simultaneously classify observations drawn from a finite mixture and estimate regression model parameters. Graca Trinadade et.al [22] introduced a new application of the Seqential Quadratic Programming algorithm to perform clustering on aggregate panel data with application to market segmentation study.

The concept of market segmentation emerged in marketing. Market Segmentation is defined as representing a heterogeneous market as a set of homogeneous submarkets. Segmentation involves creating groups of customers who show similar characteristics and can be targeted with customized strategies in context of product markets. Market segmentation is the process of segmenting a market into distinct groups of customers who share similar characteristics, needs, or behaviors. This approach enables the companies to customize their business strategies for each segment to improve the company sales and profits. [25]. Philippe Masset[24] applied market segmentation to wine data to predict the price of fine wines over their life cycle using regression approach. Tuma, M et.al [26] reviewed finite mixture models in market segmentation. Juan Prieto-Rodriguez et.al [27] investigated whether the null hypothesis of a unique segment of prices in the high end of art market can be rejected using Finite Mixture Model (FMM). Aytaç B et.al [28] considered the two regression-based techniques used to detect herding among investors. Herding is described as the tendency of investors to imitate others by suppressing their own beliefs. They also introduced an approach based on the autocorrelation of returns and tested all models on a unique dataset of wine prices. Renneboog L et.al[29] examined geographical segmentation and its effects on price formation and returns in the international art auction market. Arouri M.E et.al [30] presented a theoretical Capital Asset Pricing Model (CAPM) to price assets in different market structures and analyzed whether when markets are partially segmented using the local or the global CAPM yields significant errors in the estimation of the cost of capital for a sample of firms from developed and emerging countries. Ashish Sood et.al [31] studied a model for predicting market penetration of new products through functional regression. Carsten Hahn et.al [32] developed an approach for capturing unobserved customer heterogeneity in structural equation modeling by using a modified finite-mixture distribution approach based on partial least squares. Di Mari et. al [33] developed a two step approach to build penalized clusterwise linear regression modeling. Clusterwise linear regression models are used to build efficient strategic decision making models in the field of market analytics.

In all these papers, it was assumed that feature vector of the segmentation data set follows Gaussain distribution and the whole data set is represented by mixture of Gaussian distribution. The major drawback of the Gaussian model is, it assumes feature vector is mesokurtic. But in some data sets the feature vector associated with data may not have mesokurtic distribution. Hence, to build accurate modeling, it is necessary to generalize the Gaussian model. One of the generalization is including platy, lepty and meso kuritc distributions. Generalized Gaussian distribution is capable of describing platy, lepty and meso kurtic distributions. Very little work has been reported in the literature regarding data segmentation using mixture regression models with Generalized Gaussian Distributed errors. To develop efficient data segmentation, in this paper an algorithm is developed assuming that the feature vector associated with the data set follows a multivariate generalized Gaussian mixture model and proposed method is applied super market dataset [21] to segment customers based on product category into low profit, medium profit and high profit margin contributed customers to the super market store.

The rest of the paper is presented as follows: Section 2 is concerned with regression model with Generalized Gaussian Mixture Model. Section 3 describes the K-Means algorithm for identifying the number of clusters using regression model. Section 4 deals with the initialization of the model parameters. Section 5 provides the estimation of model parameters using Expectation and Maximization (EM) algorithm. Section 6 describes the segmentation algorithm for regression models with Generalized Gaussian Mixture model. Section 7 deals with experimental results and performance evaluation of the model. Section 8 deals with conclusions.

In this paper we follow the following notations:

$i = 1,2,3, \ldots \ldots I$ are subjects / observations/ data points.

$j = 1,2,3, \ldots \ldots J$ are independent variables.
$Y_i =$ the value of the dependent variable for subject / observation.
$x_{ij} =$ the value of the j[th] independent variable.
$k = 1,2, \ldots \ldots K$ Clusters.

2. REGRESSION MODEL WITH GENERALIZED GAUSSIAN MIXTURE MODEL

The mixture of regression models are composed using the conditional mixture and maximum likelihood methodology. The CLR models based on the maximum likelihood methodology are also called as finite mixture models for regression problems and finite mixture of linear regression [6]. Finite mixture models for regression were discussed in [7]. In the 1990s, these models were extended by mixing standard linear regression models and generalized linear models [8].

In the mixture model method, it is assumed that the data points (observations) arise from k distinct random clusters [1]. Each of the clusters is modeled by specific density function. Let z be a random variable and P(z, φk)be a probability density function for each k=1,2,.. K. Then the variable z is said to arise from a finite mixture model if it has a density function in the following represented form. $h(z, \varphi) = \sum_{k=1}^{K} \alpha_k \, P(z, \varphi k) \quad \alpha_k \geq 0, \quad \sum_{k=1}^{K} \alpha_k = 1$ (2.1)

where φk is the component parameter vector for the density function and α_k is the mixing proportion of the component k, k= 1,2, …. K.

The density function P can be used to formulate the relationship between the independent and dependent variables in the regression. Let X is independent feature vector, Y is response (dependent) feature vector of a dataset D and assume that Y is distributed as a finite mixture of conditional Generalized Gaussian densities.

The probability density function (pdf) of the Generalized Gaussian Distribution (GGD) with mean$\mu = 0$ is defined as follows.

$$f(x, \theta) = \frac{\theta k(\theta)}{2\sigma} e^{-A(\theta)\left|\frac{x}{\sigma}\right|^{\theta}} \quad (2.2)$$

Where $A(\theta) = \left(\frac{\Gamma\left(\frac{3}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)}\right)^{\frac{\theta}{2}}, k(\theta) = \frac{\Gamma\left(\frac{3}{\theta}\right)^{\frac{1}{2}}}{\Gamma\left(\frac{1}{\theta}\right)^{\frac{3}{2}}}$, σ is standard deviation, θ is shape parameter and $\Gamma(.)$ is Gamma function. Figure 2.1 represents the frequency curve of Generalized Gaussian Distribution with different shape parameters.

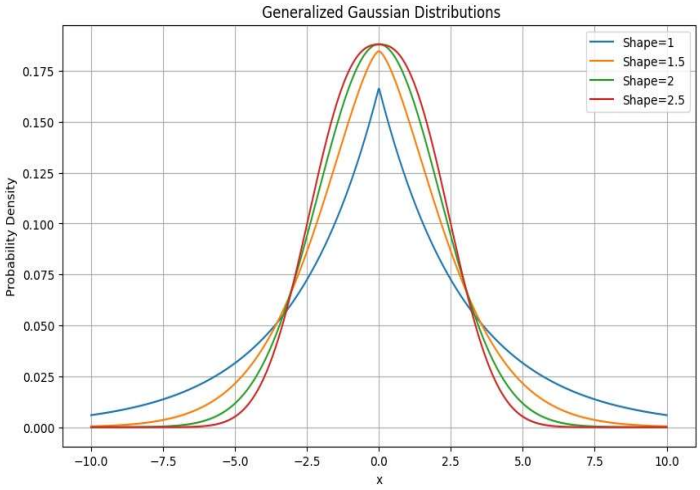

Figure 2.1: Generalized Gaussian Distributions With Different Shape Parameters

The finite mixture regression model with k components is

$h(Y|X, \varphi) = \sum_{k=1}^{K} \alpha_k \, P(Y|X, \varphi k) \quad \alpha_k \geq 0, \sum_{k=1}^{K} \alpha_k = 1$ (2.3)

where, $P(Y|X, \varphi k)$ is the probability density function of the kth component and φ is the vector of all parameters. Then CLR is modeled as a finite mixture or sum of conditional univariate densities as

$Y_i \sim \sum_{k=1}^{K} \alpha_k \, Pij(Yi|X, \varphi k)$ (2.4)

Where, P_{ij} are univariate Generalized Gaussian densities. The model becomes a mixture of standard linear regression model. If P_{ij} are members of the exponential family then we get a mixture of generalized linear regression models [9].

A mixture model based approach to regression analysis assumes that the observations of a data set originate from various groups with unknown segment affiliation.

The mixture of linear regression is defined as follows.

$Y_i = \sum_{k=1}^{K} \alpha_k f_k(Y_i|\varphi_k) + \epsilon \,, \; i = 1,2,3, .I$ (2.5)

Y_i is the dependent variable , α_k is the relative size (mixture proportion) of segment k. where $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k > 0 \; \forall \; k = 1,2, \ldots \ldots K$

Now Y_i is distributed as a finite sum or mixture of conditional univariate Generalized Gaussian Distribution (GGD).

$Y_i = \sum_{k=1}^{K} \alpha_k f_{ik}\left(y_i \middle| x_{ij}, \sigma, \beta_{ij}\right)$ Where β_{ij} is regression coefficient.

$$Y_i = \sum_{k=1}^{K} \alpha_k \frac{\theta_k k(\theta_k)}{2\sigma_k} e^{-A(\theta_k)\left(\frac{\left|Y_i - (\beta_0 + \beta_{1k} x_{1i} + \beta_{2k} x_{2i} + \ldots\ldots + \beta_{nk} x_{ni})\right|}{\sigma_k}\right)^{\theta_k}} \quad (2.6)$$

Journal of Theoretical and Applied Information Technology
30th April 2025. Vol.103. No.8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. K-MEANS ALGORITHM FOR IDENTIFYING THE NUMBER OF CLUSTERS USING REGRESSION MODEL

The K-means algorithm for obtaining the number of clusters is as follows [23]

Step 1: Identify the value of k initially

Step 2: Initialize k cluster centriods.

Step 3: Determine the cluster memberships of N observations by assigning them to the nearest cluster centroid.

Step 4: Re estimate the k cluster centroids by assuming that the membership found above is correct.

Steps 5: If none of the N observations changed membership in the last iteration, then exit; else go to step 3

4. INITIALIZATION OF MODEL PARAMETERS

The process of identifying the initial estimates of the parametric set for the given linear regression model based on GGD, one need to update the parameters using EM algorithm. The main constraint in the execution of EM algorithm is that it is totally dependent on the number of clusters and initial estimates of the model parameters [3]. The initial estimates are obtained by using method of moments and ordinary least square method.

The updated equations are to be calculated for α_k (the mixing parameter), σ_k(Standard Deviation) and β_{jk} (Regression Coefficient). Since the process is unsupervised, the initial knowledge about the parameters within the data is highly unpredictable.

5. ESTIMATION OF THE MODEL PARAMETERS USING EM ALGORITHM

In this section, we consider estimation of model parameters using Expectation Maximization (EM) algorithm that maximizes the likelihood function of the model [3]. Given a sample of I independent subjects / observations we can form the likelihood expression

$$L = \prod_{i=1}^{I}\left[\sum_{k=1}^{K}\alpha_k \frac{\theta_k k(\theta_k)}{2\sigma_k} e^{-A(\theta_k)\left(\frac{|Y_i-(\beta_0+\beta_{1k}x_{1i}+\beta_{2k}x_{2i}+\ldots\ldots+\beta_{nk}x_{ni})|}{\sigma_k}\right)^{\theta_k}}\right] \quad (5.1)$$

where $0 \le \alpha_k \le 1, \sum_{k=1}^{K}\alpha_k = 1, \sigma_k > 0$

The log likelihood function is $\ln L = \sum_{i=1}^{I} ln\left[\sum_{k=1}^{K}\alpha_k \frac{\theta_k k(\theta_k)}{2\sigma_k} e^{-A(\theta_k)\left(\frac{|Y_i-(\beta_0+\beta_{1k}x_{1i}+\beta_{2k}x_{2i}+\ldots\ldots+\beta_{nk}x_{ni})|}{\sigma_k}\right)^{\theta_k}}\right]$ (5.2)

To estimate the values of parameters $\alpha_k, \sigma_k, \beta_{jk}$, EM algorithm comprising of two steps i.e. Expectation (E) step and Maximization (M) step is utilized. The primary step in the EM algorithm needs the estimation of initial estimates from a given dataset. The refined estimates of parameters $\alpha_k, \sigma_k, \beta_{jk}$are obtained by maximizing the expected value likelihood or log likelihood. The procedure given by [4] is utilized to estimate the shape parameter θ_k.

The idea of the EM algorithm is then to iteratively calculate the maximum likelihood estimate of the unknown parameter set $\varphi = (\alpha_k, \sigma_k, \beta_{jk})$. The first step of EM algorithm is to estimate initial model parameters $\alpha_k, \sigma_k, \beta_{jk}$ from a given observations of data. The second step is to maximize $Q(\varphi, \varphi^{(1)})$ [5]. Using the steps in the EM algorithm, we get the following updated equations for the model parameters.

for α_k :

$$\alpha_k = \frac{\sum_{i=1}^{I}\widehat{p_{ik}}}{I} \quad (5.3)$$

for β_{jk}:

$$\sum_{i=1}^{I}\widehat{p_{ik}}\frac{A(\theta_k)}{\sigma_k^{\theta_k}}\theta_k\, sgn\left(Y_i-\left(\beta_0+\sum_{i=1,j=1,k=1}^{i=I,j=J,k=K}\beta_{jk}x_{ij}\right)\right)\left|Y_i-\left(\beta_0+\sum_{i=1,j=1,k=1}^{i=I,j=J,k=K}\beta_{jk}x_{ij}\right)\right|^{\theta_k-1}x_{ij}=0 \quad (5.4)$$

As a special case if $\theta_k = 2$ we have Gaussian distribution. Then for $\theta_k = 2$ we have

$$\sum_{i=1}^{I}\widehat{p_{ik}}\left|Y_i-\left(\beta_0+\sum_{i=1,j=1,k=1}^{i=I,j=J,k=K}\beta_{jk}x_{ij}\right)\right|x_{ij}=0$$

for σ_k: $\sum_{i=1}^{I}\frac{\widehat{p_{ik}}}{\sigma_k}\left(\theta_k A(\theta_k)\left|Y_i-\left(\beta_0+\sum_{i=1,j=1,k=1}^{i=I,j=J,k=K}\beta_{jk}x_{ij}\right)\right|^{\theta_k}\sigma_k^{-\theta_k}-1\right)=0$ (5.5)

As a special case if $\theta_k = 2$ we have Gaussian distribution. Then for $\theta_k = 2$ we have

$$\sigma_k = \left(\frac{\sum_{i=1}^{I}\widehat{p_{ik}}\left(Y_i-\left(\beta_0+\sum_{i=1,j=1,k=1}^{i=I,j=J,k=K}\beta_{jk}x_{ij}\right)\right)^2}{\sum_{i=1}^{I}\widehat{p_{ik}}}\right)^{\frac{1}{2}} \quad (5.6)$$

Solving the equations (5.3),(5.4) and (5.5) simultaneously and iteratively , the refined estimates of the model parameters $\alpha_k, \sigma_k, \beta_{jk}$ can be obtained.

Once estimates of $\alpha_k, \sigma_k, \beta_{jk}$ are obtained, one can assign each observation i to each cluster k via the estimated posterior probability using Bayes rule.

$$\widehat{p_{ik}} = \frac{\widehat{\alpha_k} f_{ik}\left(Y_i \middle| x_{ij}, \widehat{\alpha_k}, \widehat{\beta_{ik}}\right)}{\sum_{k=1}^{K} \widehat{\alpha_k} f_{ik}\left(Y_i \middle| x_{ij}, \widehat{\alpha_k}, \widehat{\beta_{ik}}\right)} \quad (5.7)$$

Assign observation i to cluster k iff $\widehat{p_{ik}} > \widehat{p_{il}} \forall\, l \neq k = 1,2, \ldots \ldots K$.

5.1 Expectation – Maximization Algorithm for GGD Error Regression Model

Step1: Select the initial parameters

Step 2: Obtain revised estimates of the parameters $\alpha_k, \sigma_k, \beta_{jk}$ using equations (5.3), (5.4) , (5.5) and (5.6).

Step 3: Repeat the process until the parameters do not change or the difference in successive computations is within the given threshold value.

Step 4: Write the final estimates of parameters $\alpha_k, \sigma_k, \beta_{jk}$

6. SEGMENTATION ALGORITHM FOR REGRESSION MODELS WITH GENERALIZED GAUSSIAN MIXTURE MODEL

In this section, the segmentation algorithm for regression models with Generalized Gaussian Distribution is presented for identifying the new data tuple with one of the available clusters. The steps involved in this algorithm are as follows:

Step 1: Draw the scatter surface diagram for the training data set in order to obtain the initial number of clusters by using the k-means algorithm.

Step 2: Obtain initial estimates of the model parameters.

Step 3: Obtain the refined estimates of the model parameters using the updated equations of the EM algorithm given in section 4.

Step 4: For a new data point, compute the conditional likelihood with the model parameters of the k^{th} class and assign it to the segment for which the sample conditional likelihood is maximum. i.e. the classification is $C=\text{argmax}_k P(D_t \mid C_k)$. where C is the maximum likelihood class and D_t is the new data point.

7. EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION

In this section, the utility of the developed algorithm for segmenting marketing data is demonstrated. The dataset was collected from Kaggle dataset repository [21]. This dataset has 21 features among them category (category of the product ordered), Sales (Sales of the Product), Quantity (Quantity of the Product) and Profit (Profit/Loss incurred) are considered as relevant features for this study. Here there are three groups of product categories like Office Supplies, Furniture and Technology. After considering super market dataset, it was understood that two features sales(X_1) and quantity(X_2) are most important for determining the profit(Y) margins such as low profit margin, medium profit margin and high profit margin of the store. Here Office Supplies category products falls under low profit margin, Furniture category products comes under medium profit margin and Technology category products under high profit margin. To identify the margins of the profit, it is required to segment the data set into various clusters based on sales and quantity variables. The number of clusters in the super market data is not known and requires unsupervised learning algorithms to identify various margins of profit. Hence a study is carried out by collecting a sample of 80 data points with sales and quantity variables of super market data set.

Using k-means algorithm, the number of profit margins according to sales and quantity is determined. For implementing the k-means algorithm, the initial number of clusters is required. Hence, using the training data, the sample data points are plotted in scatter responses through a 3-dimensional graph shown in figure 7.1.

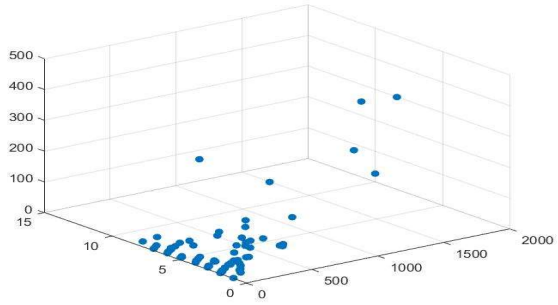

Figure 7.1: Scatter plot of data points

Using the initialization of parameters discussed in section 4, the initial estimates of parameters $\alpha_k, \sigma_k, \beta_{jk}$ are obtained for 3 profit margins such that low profit margin corresponds to cluster 1, medium profit margin corresponds to cluster 2 and high profit margin corresponds to cluster 3. The computed initial estimates of the model parameters are presented in Table 7.1

Table 7.1: Initial Estimates Of The Model Parameters

Parameter	Cluster 1(Low Profit Margin)	Cluster 2 (Medium Profit Margin)	Cluster 3(High Profit Margin)
α_k	0.9000	0.0500	0.0500
σ_k	313.5360	932.7305	12895.5578
β_{jk} Intercept	17.0813	55.4252	193.5787
Coefficient 1	0.1270	0.0504	0.2853
Coefficient 2	-1.6845	-2.6176	-63.5386

Using these initial estimates and the EM algorithm, the refined estimates of parameters for each cluster are obtained and presented in Table 7.2.

Table 7.2: Final Estimates Of The Model Parameters

Parameter	Cluster 1(Low Profit Margin)	Cluster 2 (Medium Profit Margin)	Cluster 3(High Profit Margin)
α_k	0.7277	0.2052	0.0671
σ_k	52.7779	190.2404	2426.3718
β_{jk} Intercept	10.4305	64.2997	230.1186
Coefficient 1	0.0863	0.0853	0.2934
Coefficient 2	-0.5878	-8.3996	-63.2065

With these final estimates, the 3 clusters of profit margins are estimated as

Cluster 1: Office Supplies Product category (low profit margin)

The estimated regression equation is:

$Y= 10.4305 + 0.0863X_1 – 0.5878X_2$

Here X_1 represents sales, X_2 represents quantity and Y represents profit

Cluster 2: Furniture Product Category (medium profit margin)

The estimated regression equation is:

$Y= 64.2997 + 0.0853X_1 – 8.3996X_2$

Cluster 3: Technology Product Category (high profit margin)

The estimated regression equation is:

$Y= 230.1186 + 0.2934X_1 – 63.2065X_2$

. Therefore, the model characterizes the whole data set is a three-component mixture of Generalized Gaussian Mixture Model (GGMM) whose component weights are: α_1 =0.7277, α_2=0.2052, α_3=0.0671, respectively. For evaluating the developed algorithm, the test data consisting of 80 data points is considered. The developed unsupervised algorithm using GGMM identified 58 data points as low profit margin, 16 data points as medium profit margin and 6 data points as high profit margin. For evaluating the performance of the proposed algorithm, accuracy, misclassification rate, precision, recall and F-measure are used. For the proposed unsupervised learning algorithm of GGMM, the performance measures for each cluster are computed and presented in Table 7.3.

Table 7.3: Performance Measures Of The Mixture Of GGMM Classifier

	True Positive Rate(TPR) Recall	Precision	False Discovery Rate	F-Measure
Cluster 1	0.9682	0.9839	0.0317	0.9760
Cluster 2	0.9166	0.7857	0.0833	0.8461
Cluster 3	0.8000	1.0000	0.2000	0.8888

To compare the efficiency of the developed GGMM classifier with earlier Gaussian Mixture Model (GMM) classifier, recall, precision, false discovery rate and F-measure are computed and presented in Table 7.4.

Table 7.4: Performance Measures Of The Mixture Of GMM Classifier

	True Positive Rate(TPR) Recall	Precision	False Discovery Rate	F-Measure
Cluster 1	0.9365	0.9672	.0635	0.9516
Cluster 2	0.9166	0.7857	0.0833	0.8461
Cluster 3	0.8000	1.0000	0.2000	0.8888

Comparing Table 7.3, Table 7.4 it is observed that the F value for cluster 1 using the proposed classifier is more compared to that of the classifier with GMM. The f value for cluster 2 and cluster 3 are equal in both proposed classifier and classifier with GMM.

To compare the efficiency of the developed unsupervised algorithm with existing unsupervised learning algorithm with GMM model for both sales and quantity variables, the same test data were considered and the accuracy and misclassification rates were computed. Table 7.5 presents the accuracy and misclassification rates of GGMM and GMM classifiers.

Table 7.5: Performance evaluation of accuracy & Misclassification rate

Classifier with	Accuracy	Misclassification rate
GGMM	0.9500	0.0500
GMM	0.9250	0.0750

From Table 7.5 it is observed that the accuracy of GGMM classifier is more compared to the accuracy of GMM classifier and the misclassification rate of GGMM classifier is lesser compared to the misclassification rate of GMM classifier.

The other parameters of both GGMM and GMM classifiers are presented in Table 7.6 and Table 7.7

Table 7.6: The Other Parameters Of Both GGMM And GGM Classifiers

Classifier with	Shape	No.of iterations	Log likelihood
GGMM	1.3760	13	-332.8420
GMM	2	11	-333.0135

Table 7.7: The Root Mean Square Error (RMSE) Of Different Clusters

Classifier with	Cluster 1	Cluster 2	Cluster 3
GGMM	1.8825	6.8082	45.1556
GMM	1.6710	7.9293	54.7316

8. CONCLUSIONS

This paper deals with the development and analysis of a novel method in segmentation algorithm for market analytics using mixture regression models with Generalized Gaussian Distribution. In market segmentation so far the algorithms developed using mixture regression models with Gaussian distribution only. For the first time we developed an unsupervised learning algorithm for market segmentation using Generalized Gaussian mixture regression models under Bayesian framework. This algorithm is more suitable for analyzing all types of data sets that exhibit behaviours such mesokurtic, playkurtic and leptokurtic. This algorithm is applied for analyzing the realistic situations in market analytics, business analytics, financial analytics, HR analytics etc. where the variables under study are correlated and follows Generalized Gaussian Distribution.

Another important feature of this developed algorithm is integration of k-means with model based method in learning algorithms. The learning algorithm is developed based on component maximum likelihood under Bayesian framework. Hence, it is assumed that the feature vector is generated from a heterogeneous population which can be characterized by a finite mixture of regression models with Generalized Gaussian Distribution. The model parameters are estimated using EM algorithm.

The performance of the proposed algorithm is evaluated using the super market data set. The experimental results revealed that the proposed algorithm outperforms the existing learning algorithms. This learning algorithm can be extended to the integration of hierarchical clustering algorithms with mixture regression models using Generalized Gaussian Distribution.

REFERENCES

[1] Wayne S. DeSarbo and William L. Cron .(1988), A maximum likelihood methodology for clusterwise linear regression, Journal of Classification, 5, pp. 249-282.

[2] Qiang Long, Adil Bagirov, Sona Taheri , Nargiz Sultanova , and Xue Wu. (2023), Methods and Applications of Clusterwise Linear Regression: A Survey and Comparison, ACM Trans.Knowl.Discov. Data, 17, 3, pp 1-54.

[3] Mclanchlan, G. and Peel. D. (2000), The EM Algorithm for parameter estimation, John Wiley and Sons New York.

[4] Shaoquan YU, Anyi Zhang, Hongwei LI (2012), A Review on estimating the Shape Parameter of Generalized Gaussian Distribution , Journal of Information Systems, Volume 8(21), pp. 9055-9064.

[5] Bilmes. Jeff A. (1998), A Gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Intl. Computer Science Institute, Berkely.

[6] Susana Faria and Gilda Soromenho (2010), Fitting mixtures of linear regressions, Journal of Statistical Computation and Simulation 80, 2 pp. 201-225.

[7] Richard E. Quandit(1972), A new approach to estimating switching regressions. Journal of American Statistical Association 67, 338 pp. 306-310.

[8] Michel Wedel and Wayne S. DeSarbo (1995), A mixture likelihood approach for generalized linear Models, Journal of Classification 12, 1, pp. 21-55.

[9] Michel Wedel and Cor Kistemaker(1989), Consumer benefit segmentation using clusterwise linear regression. International

Journal of Research in Marketing 6, 1, pp. 45-59.

[10] Christian Preda and Gilbert Saporta (2005), Clusterwise PLS regression on a stochastic process, Computational Statistics & Data Analysis 49, 1, pp. 99-108.

[11] Jagadeesh P. Ganjigatti, Dilip K. Pratihar and A.Roy Choudhury (2007), Global versus cluster-wise regression analyses for prediction of bead geometry in MIG welding process, Journal of Materials Processing Technology 189, 1- 3, pp. 352-366.

[12] Mukesh Khadka and Alexander Paz(2017), Comprehensive clusterwise linear regression for pavement management systems, Journal of Transportation Engineering, Part B : Pavments 143, 4 , 04017014.

[13] Adil M. Bagirov, Arshad Mahmood and Andrew Barton (2017), Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach, Atmospheric Research 188, pp. 20-29.

[14] Jean – Michel Poggi and Bruno Portier (2011), forecasting using clusterwise linear regression. Atmospheric Environment 45, 38, pp. 7005-7014.

[15] Ye Chow Kuang , Melanie Ooi(2024), Performance Characterization of Clusterwise Linear Regression Algorithms, Wiley Interdicplinary Reviews: Computational Statistics, pp. 1-16.

[16] Yifan Zhang, Ducan K.H. Fong , Wayne S.DeSarbo(2021), A Generalized Ordinal Finite Mixture Regression Model Market Segmentation, International Journal of Research in Marketing.

[17] Ting Li, Xinyuan Song, Yingying Zhang, Hongtu Zhu, Zhongyi Zhu(2021), Clusterwise Functional Linear Regression Models , Computational Statistics and Data Analysis , pp. 1-15.

[18] Kaisa Joki , Adil M. Bagirov, Napsu Karmitsa, Marko M. Makela, Sona Taheri(2020), Clusterwise Support Vector Linear Regression, European Journal of Operational Research, pp. 19-35.

[19] Paul W.Murray, Bruno Agard, Marco A. Barajas(2017) ,Market Segmentation through data mining: A Method to extract behaviors from a noisy data set, Computers & Industrial Engineering , pp. 233-252.

[20] Cathy W.S. Chen, Jennifer S.K. Chan, Mike K.P. So, Kevin K.M. Lee (2011), Classification in Segmented regression problems, Computational Statistics and Data Analysis, pp. 2276-2287.

[21]https://www.kaggle.com/datasets/vivek468/superstore-dataset-final.

[22] Graca Trindade, Jose G. Dias, Jorge Ambrosio (2017), Extarcting clusters from aggregate panel data: A Market segmentation study, Journal of Applied Mathematics and Computation, 296, pp. 277-288.

[23] K.Vedavathi , K,Srinivasa Rao, K.Nirupama Devi (2014), Unsupervised learning algorithm for time series using bivariate AR(1) model, Expert Systems with Applications, 41, 3402-3408.

[24] Philippe Masset(2024), Market segments and pricing of fine wines over their lifecycle, Economic Modelling,141,106915.

[25] Wedel, M.Kamakura, W.A (2000) , Market Segmentation: Conceptual and Methodological Foundations, Second Edition.

[26] Tuma, M and Decker, R(2013) , Finite Mixture Models in Market Segmentation: A Review and Suggestions for Best Practices, The Electronic Journal of Business Research Methods 11,1, pp 02-15.

[27] Juan Prieto-Rodriguez, Marilena Vecco(2021), Reading between the lines in the art market: Lack of transparency and price heterogeneity as an indicator of multiple equilibria, Economic Modelling, 102, 105587.

[28] Aytaç B, Coqueret G, Mandou C (2018), Herding behavior among wine investors, Economics Modelling, 68, pp.318-328

[29] Renneboog L, Spaenjers C (2014), Investment returns and economic fundamentals in international art markets, in: Canvases and Careers in a Cosmopolitan Culture. On the Globalization of Contemporary Art Markets, O. Velthuis and S. Baia-Curioni (eds.), Oxford University Press.

[30] Arouri M.E, Rault C, Sova R, Sova A(2013), Market Structure and the Cost of Capital, Economics Modelling, 31, pp.664-671.

[31] Ashish Sood, Gareth M. James, Gerard J. Tellis, (2008) Functional Regression: A New Model for Predicting Market Penetration of New Products. Marketing Science 28(1) , pp.36-51.

[32] Carsten Hahn,Michael D. Johnson,Andreas Herrmann, Frank Huber(2002), Capturing Customer Heterogeneity Using A Finite Mixture PLS Approach, Schmalenbach Business Review 54, pp.243 – 269.

[33] Di Mari, R., Rocci, R., & Gattone, S. A. (2023). LASSO–penalized clusterwise linear regression modelling: a two–step approach. *Journal of Statistical Computation and Simulation*, *93*(18), 3235–3258. https://doi.org/10.1080/00949655.2023.2220058