International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

Gaussian Mixture Model Fitting Using Differential Linear Regression

Naleli Jubert Matjelo[1], Molise Mokhomo[2]

[1]National University of Lesotho, Department of Physics and Electronics, P. O. Roma 180, Lesotho
[2]Mokuru, Cape Town, Western Cape, South Africa.

Abstract – *In this paper, we explore the possibility of formulating the Gaussian mixture model (GMM) fitting problem as a linear regression problem based on the differential relationship in the GMM data points. This approach is not popular in the literature yet it is very powerful in simplifying and transforming many nonlinear regression problems into linear regression problems. We demonstrate here with fitting a dataset to a two-component GMM, something which would commonly be considered as a nonlinear regression problem. In this differential formulation, the approach is shown to be sensitive to noise and an autoregressive formulation approach is suggested to avoid noise amplification.*

Key Words: **Gaussian Mixture Model, Regression, Noise Amplification, Parameter Estimation, Model Fitting.**

1. INTRODUCTION

Parameter estimation problem often comes up whenever some experimental data has to be fitted to the model of choice. This could be for various reasons including system identification & characterization, behavioral analysis, model-based control & state estimation, as well as forecasting, smoothing, and filtering [1], [2], [3]. One of the models used to estimate almost any smooth continuous function is the Gaussian mixture model (GMM), composed of the weighted average of Gaussian functions [4]. Two examples of GMM applications are classification & clustering [5] and nonGaussian probability distribution estimation [6]. The properties of Gaussian functions (including self-conjugacy, stationarity, local support, etc) make them very attractive for Bayesian evolution [7], and as the finite basis set for estimating other complicated functions [8]. With GMM being a nonlinear model, fitting data to such a model is often formulated as nonlinear regression. In this formulation there exists no closed-form solution to the problem hence the iterative optimization-based algorithms are adopted to solve this problem [9].

In this paper, we adopt a differential formulation of the regression problem which allows the model parameter estimation problem to be transformed from a nonlinear regression to a linear regression problem. The approach is based on shifting the problem from how the model output is related to its independent variables to how it relates to its derivatives. We consider a two-component GMM to demonstrate the differential-based linear regression formulation of the GMM fitting problem.

The rest of this paper is organized as follows. Section 2 presents a two-component GMM and the resulting differential linear regression problem formulation. Section 3 presents model-fitting simulation results and discussion. Section 4 concludes this work with a summary of major findings and some remarks.

2. LINEAR REGRESSION MODEL FOR GMM

2.1 Gaussian Mixture Model

A general one-dimensional M-component GMM $y(x)$ can be represented as follows,

$$y(x) = \sum_{i=1}^{M} a_i e^{-b_i(x-\mu_i)^2} \quad (1)$$

with a_i as the i^{th} component weight, b_i being inversely related to the i^{th} component variance and μ_i being the mean of the i^{th} component. In this paper, we will restrict ourselves to $M = 2$ however, the same concept applies to higher values of M. With $M = 2$ we have the following model with six unknown parameters to be determined from the data,

$$y(x) = a_1 e^{-b_1(x-\mu_1)^2} + a_2 e^{-b_2(x-\mu_2)^2} \quad (2)$$

2.2 Gaussian Mixture Model In Differential Form

Differentiating equation (*2*) twice and relating the equation (*2*) with its first two derivatives we obtain the following differential equation,

$$C_2(x)y'' + C_1(x)y' + C_0(x)y = 0 \quad (3)$$

with variable coefficients $C_n(x)$ given by,

$$C_2(x) = c_1x + c_2 \quad (4)$$

$$C_1(x) = c_3x^2 + c_4x + c_5 \quad (5)$$

$$C_0(x) = c_6x^3 + c_7x^2 + c_8x + c_9 \quad (6)$$

where the constant coefficients c_m are given by,

$$c_1 = b_1 - b_2 \quad (7)$$

$$c_2 = b_2\mu_2 - b_1\mu_1 \quad (8)$$

$$c_3 = 2(b_1^2 - b_2^2) \quad (9)$$

$$c_4 = 4(b_2^2\mu_2 - b_1^2\mu_1) \quad (10)$$

$$c_5 = 2(b_1^2\mu_1^2 - b_2^2\mu_2^2) + b_2 - b_1 \quad (11)$$

$$c_6 = 4b_1b_2(b_1 - b_2) \quad (12)$$

$$c_7 = 4b_1b_2\big((2b_2 - b_1)\mu_2 + (b_2 - 2b_1)\mu_1\big) \quad (13)$$

$$c_8 = 4b_1b_2\big((b_1 - b_2)\mu_1\mu_2 + b_1\mu_1^2 - b_2\mu_2^2\big) \quad (14)$$

$$c_9 = 2b_1b_2(2\mu_1\mu_2(b_2\mu_2 - b_1\mu_1) + \mu_2 - \mu_1) \quad (15)$$

In the next section, we show, without going into detail, the autoregressive representation of the differential model in equation (3).

2.3 Gaussian Mixture Model In Autoregressive Form

Equation (3) above can be discretized, with a discretization parameter (or sampling interval) h in x, to give the following difference equation (or an autoregressive model),

$$y_{i+1} = \frac{2C_2(x_i) - h^2C_0(x_i)}{hC_1(x_i) - C_2(x_i)}\, y_i + \frac{hC_1(x_i) - C_2(x_i)}{hC_1(x_i) - C_2(x_i)}\, y_{i-1} \quad (16)$$

with i indicating the i^{th} sample or discrete data point index. It was shown in [10], [11] that the autoregressive formulation is less prone to noise than the differential formulation due since the differential operation (i.e. equation (3)) amplifies high-frequency noise while the integration operation (i.e. equation (16)) is a low pass filter. From [10], [11] it is made clear how one can switch between the two representations when formulating the linear regression model hence in this paper, we will only focus on one representation and that being the linear regression formulation based on equation (3). In the next section, we present the linear regression problems based on a given data set and show the solution.

2.4 Differential Linear Regression Model

Given a data $y_i(x_i)$ of size N (i.e. $i = 1, 2, 3, \dots, N$) to fit a GMM, we can formulate the linear regression problem using equations (3-15) as shown in the cost function below,

$$J(\alpha_n) = \sum_{i=1}^{N}\begin{pmatrix}\alpha_1 x_i y_i'' + \alpha_3 x_i^2 y_i' + \alpha_4 x_i y_i' + \alpha_5 y_i' + \\ \alpha_6 x_i^3 y_i + \alpha_7 x_i^2 y_i + \alpha_8 x_i y_i + \alpha_9 y_i + y_i''\end{pmatrix}^2 \quad (17)$$

with $\alpha_i = c_i c_2^{-1}$. The resulting solution to this linear regression problem is given by the following matrix equation,

$$\begin{bmatrix}\alpha_1\\ \alpha_3\\ \alpha_4\\ \alpha_5\\ \alpha_6\\ \alpha_7\\ \alpha_8\\ \alpha_9\end{bmatrix} = -\sum_{i-1}^{N}\begin{bmatrix} x_i^2y_i''^2 & x_i^3y_i'y_i'' & x_i^2y_i'y_i'' & x_iy_i'y_i'' & x_i^4y_iy_i'' \\ x_i^3y_i'y_i'' & x_i^4y_i'^2 & x_i^3y_i'^2 & x_i^2y_i'^2 & x_i^5y_iy_i' \\ x_i^2y_i'y_i'' & x_i^3y_i'^2 & x_i^2y_i'^2 & x_iy_i'^2 & x_i^4y_iy_i' \\ x_iy_i'y_i'' & x_i^2y_i'^2 & x_iy_i'^2 & y_i'^2 & x_i^3y_iy_i' \\ x_i^4y_iy_i'' & x_i^5y_iy_i' & x_i^4y_iy_i' & x_i^3y_iy_i' & x_i^6y_i^2 \\ x_i^3y_iy_i'' & x_i^4y_iy_i' & x_i^3y_iy_i' & x_i^2y_iy_i' & x_i^5y_i^2 \\ x_i^2y_iy_i'' & x_i^3y_iy_i' & x_i^2y_iy_i' & x_iy_iy_i' & x_i^4y_i^2 \\ x_iy_iy_i'' & x_i^2y_iy_i' & x_iy_iy_i' & y_iy_i' & x_i^3y_i^2 \end{bmatrix}$$

$$\begin{bmatrix} x_i^3y_iy_i'' & x_i^2y_iy_i'' & x_iy_iy_i'' \\ x_i^4y_iy_i' & x_i^3y_iy_i' & x_i^2y_iy_i' \\ x_i^3y_iy_i' & x_i^2y_iy_i' & x_iy_iy_i' \\ x_i^2y_iy_i' & x_iy_iy_i' & y_iy_i' \\ x_i^5y_i^2 & x_i^4y_i^2 & x_i^3y_i^2 \\ x_i^4y_i^2 & x_i^3y_i^2 & x_i^2y_i^2 \\ x_i^3y_i^2 & x_i^2y_i^2 & x_iy_i^2 \\ x_i^2y_i^2 & x_iy_i^2 & y_i^2 \end{bmatrix}^{-1} \sum_{i-1}^{N}\begin{bmatrix} x_iy_i''^2 \\ x_i^2y_i'y_i'' \\ x_iy_i'y_i'' \\ y_i'y_i'' \\ x_i^3y_iy_i'' \\ x_i^2y_iy_i'' \\ x_iy_iy_i'' \\ y_iy_i'' \end{bmatrix} \quad (18)$$

from which the parameters of interest b_1, b_2, μ_1 and μ_2 are obtained using the following relations,

$$b_1 = \frac{\alpha_3 \pm \sqrt{\alpha_3^2 - 4\alpha_1\alpha_6}}{4\alpha_1} \quad (19)$$

$$b_2 = \frac{\alpha_3 \mp \sqrt{\alpha_3^2 - 4\alpha_1\alpha_6}}{4\alpha_1} \quad (20)$$

$$\mu_1 = \frac{8b_1b_2 - 4b_1^2 - \alpha_7}{4\alpha_1b_1(b_1 + b_2)} \quad (21)$$

$$\mu_2 = \frac{8b_1b_2 - 4b_2^2 - \alpha_7}{4\alpha_1b_2(b_1 + b_2)} \quad (22)$$

Given the solutions of these four parameters, we continue to solve for the weights a_1 and a_2 from the following linear regression constructed based on equation (2),

$$J(a_m) = \sum_{i=1}^{N}\left(\mathrm{a_1e^{-b_1(x_i-\mu_1)^2} + a_2e^{-b_2(x_i-\mu_2)^2} - y_i}\right)^2 \quad (23)$$

which leads to the following solution for two weights a_1 and a_2,

$$\begin{bmatrix}a_1\\ a_2\end{bmatrix} = \sum_{i=1}^{N}\begin{bmatrix} \mathrm{e^{-2b_1(x_i-\mu_1)^2}} \\ \mathrm{e^{-b_1(x_i-\mu_1)^2}e^{-b_2(x_i-\mu_2)^2}} \end{bmatrix}$$

$$\begin{bmatrix} e^{-b_1(x_i-\mu_1)^2}e^{-b_2(x_i-\mu_2)^2} \\ e^{-2b_2(x_i-\mu_2)^2} \end{bmatrix}^{-1} \sum_{i=1}^{N}\begin{bmatrix} y_ie^{-b_1(x_i-\mu_1)^2} \\ y_ie^{-b_2(x_i-\mu_2)^2} \end{bmatrix} \quad (24)$$

The next section presents simulation results for fitting this GMM using this linear regression approach.

4. SIMULATION RESULTS & DISCUSSION

This section presents the simulation of the proposed linear regression model for a case of noiseless data. The data points were generated by a two-component GMM shown in equation (2) with parameters $a_1 = 6.50$, $a_2 = -8.10$, $b_1 = 1.50$, $b_2 = 0.75$, $\mu_1 = 3.80$, $\mu_2 = 6.75$ and discretization parameter $h = 10^{-4}$. We then subjected this generated data to the linear regression outlined in the previous section to obtain the parameter estimates. Fig. 1 below shows the plot of the generated data and the GMM model estimation based on the linear regression.

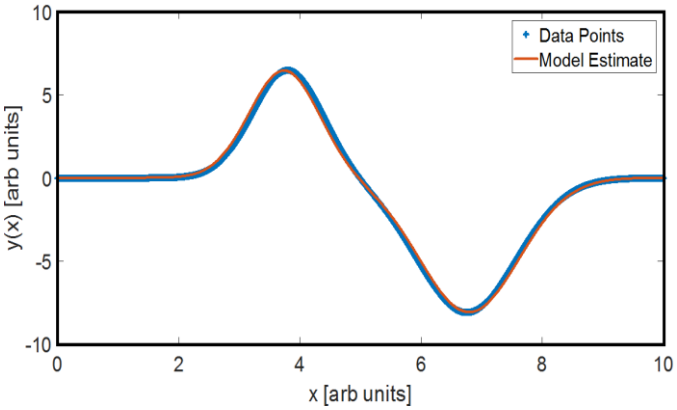

Fig. 1 *GMM fitting when discretization is* $h = 10^{-4}$.

The parameters estimated from this data using the linear regression model when the discretization parameter is set to $h = 10^{-4}$ are shown in Table 1 below.

Table 1: *Estimated parameters for discretization* $h = 10^{-4}$.

Parameter	Exact Value	EstimatedValue	% Error
a_1	6.50	6.4692	0.47%
a_2	−8.10	−8.0821	0.22%
b_1	1.50	1.4986	0.09%
b_2	0.75	0.7493	0.09%
μ_1	3.80	3.7422	1.52%
μ_2	6.75	6.7955	0.67%

These errors are mostly due to discretization as will be shown from Fig. 3 how discretization parameter size affects the estimation. Fig. 2 below shows the squared error between the data points and the model-generated points based on these estimated parameters.

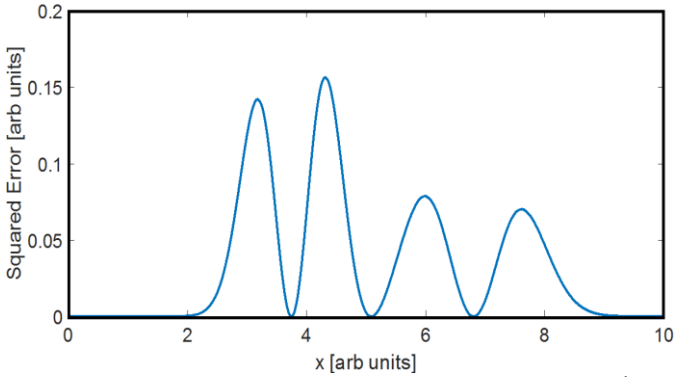

Fig. 2 *Error between data and model when* $h = 10^{-4}$.

Overall the errors seem acceptable for practical estimation purposes. In [10], [11] it was shown that this differential formulation of the linear regression is sensitive to noise. Here we wish to show that this formulation is also sensitive to discretization parameter size h. Fig. 3 below shows the same generated data plotted together with the model estimate when the discretization parameter h has been increased by a factor of 10 to $h = 10^{-3}$.

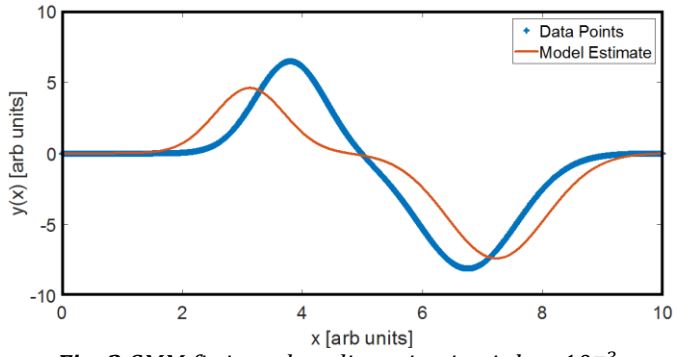

Fig. 3 *GMM fitting when discretization is* $h = 10^{-3}$.

Table 2 below shows the parameters estimated when the discretization parameter is set to $h = 10^{-4}$.

Table 2: *Estimated parameters for discretization* $h = 10^{-3}$.

Parameter	Exact Value	Estimated Value	% Error
a_1	6.50	4.5971	29.28%
a_2	−8.10	−7.3953	8.70%
b_1	1.50	1.4812	1.25%
b_2	0.75	0.7416	1.12%
μ_1	3.80	3.1238	17.79%
μ_2	6.75	7.2309	7.12%

Comparing the errors from Table 1 with those in Table 2, it is evident that the differential linear regression formulation is sensitive to discretization size. Fig. 2 below shows the squared error between the data points the graph points generated based on these estimated parameters.

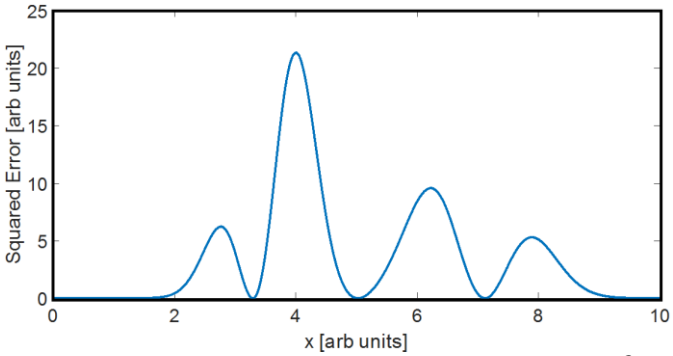

Fig. 4 Error between data and model when $h = 10^{-3}$.

Discretization size and noise aside, the differential linear regression formulation is a promising alternative to the nonlinear regression formulation for the same underlying parameter estimation problem. As shown in [10], [11] that one way to improve the differential linear regression model against noise is to adopt the autoregressive formulation approach, it would be interesting to investigate if the autoregressive formulation approach can also be more robust than the differential formulation against discretization parameter size.

5. CONCLUSIONS

In this work, we have successfully shown how the two-component GMM data-fitting problem can be presented as a linear regression problem with a closed-form form solution. The formulation gave good results, however, it was observed that the formulation is sensitive to the discretization size (or sampling frequency). The autoregressive formulation of the same problem was presented as an alternative in theory but not tested experimentally with simulated data. As part of future work, this autoregressive formulation can be explored and applied to other interesting nonlinear regression problems other than the ones involving GMMs.

REFERENCES

[1] Beck M.B. Applications of System Identification and Parameter Estimation in Water Quality Modeling. (Proceedings of the Oxford Symposium): IAHS-AISH Publ, no. 129, 1980.

[2] Ding R., Zhuang L. Parameter and State Estimator for State Space Models. The Scientific World Journal, vol. 2014, Article ID 106505, 10 pages, 2014.

[3] Luengo D., Martino L., Bugallo M., Elvira V., Sarkka S. A Survey of Monte Carlo Methods for Parameter Estimation. EURASIP J. Adv. Signal Process, 25, 2020,

[4] Reynolds D. Gaussian Mixture Models. In: Li S.Z., Jain A. (eds) Encyclopedia of Biometrics. Springer, Boston, MA, 2009.

[5] Wang Z., da Cunha C., Ritou M., Furet B. Comparison of K-means and GMM methods for contextual clustering in HSM. Procedia Manufacturing, Elsevier, 2019, 7th International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV2018), 28, pp.154-159, 2018.

[6] Tsukakoshi K., Ida K. Gaussian Mixture Distribution Analysis as Estimation of Probability Density Function and It's the Periphery, The 2015 International Conference on Soft Computing and Software Engineering (SCSE 2015), 370-377, 2015.

[7] Kunkel D., M.S. Anchored Bayesian Gaussian Mixture Models, Ph.D. Thesis, The Ohio State University, 2018

[8] Dalal, S. R., and W. J. Hall. "Approximating Priors by Mixtures of Natural Conjugate Priors." Journal of the Royal Statistical Society. Series B (Methodological), vol. 45, no. 2, pp. 278–286, 1983

[9] Ghojogh B., Ghojogh A., Crowley M., Karray F. Fitting A Mixture Distribution to Data: Tutorial, arXiv:1901.06708v2 [stat.OT] 11 Oct 2020

[10] Matjelo N.J. Differential Linear Regression Model For Frequency Estimation. International Research Journal of Engineering and Technology (IRJET), 08(07):1-7, 2021.

[11] Matjelo N.J. Autoregressive Linear Regression Model For Frequency Estimation. International Research Journal of Engineering and Technology (IRJET), 08(07):243-248, 2021.