Interactive Toolbox for two-dimensional Gaussian Mixture Modeling

Michael C.Thrun[1][0000-0001-9542-5543] , Quirin Stier[1] and Alfred Ultsch[1]

[1] Mathematics and Computer Science, Philipps-Universität Marburg, Hans-Meerwein-Straße 6, D-35032 Marburg, mthrun@informatik.uni-marburg.de

Abstract. Research data obtained during economics or human studies experiments often displays a complex distribution. Even in the two-dimensional case, the statistical identification of subgroups in research data poses an analytical challenge. Here we introduce an interactive R-based tool called "AdaptGauss2D". It enables a valid identification of a meaningful multimodal structure in two-dimensional data. With a human-in-the-loop approach, a Gaussian mixture model (GMM) can be fitted to the data. The interactive interface allows a supervised selection of the number and parameters of the GMM based on various visualizations. Integrating a Human-in-the-loop into the process of modeling two-dimensional gaussian mixtures enables the expectation-maximization (EM) algorithm to adapt to more complex GMM compared to the standard non-interactive approach. The work demonstrates that the interactive modeling process for GMM improves the quality of the model in contrast to non-interactive modeling. The improvement is shown using the datasets of EngyTime and a large flow cytometry dataset. The R package "AdaptGauss2D" is available on GitHub https://github.com/Mthrun/AdaptGauss2D.

Keywords: Gaussian Mixtures, Human-in-the-loop, Interactive ML.

1 Introduction

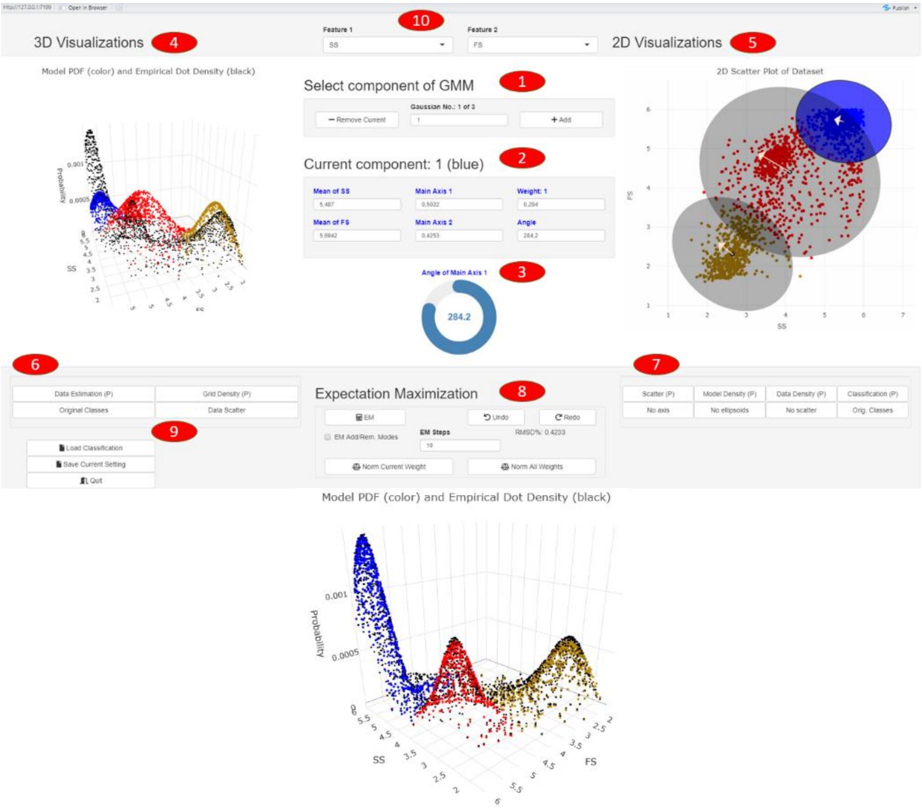
A Gaussian mixture model (GMM) is a probabilistic model that explains the chance of detecting an event x with probability p with the assumption that underlying data is generated using the weighted sum of a finite number k of normal distributions $\mathrm{N}(X|M_i, S_i)$ also known as modes or components, with means $M_i$and covariance S_i. The weighting w_i determines the relative contribution of each of these normal distributions to the mixture and is the prior probability of occurrence of the modes with $\sum_{i=1}^{k} w_i = 1$. In the two-dimensional case, a k-modal GMM is defined as $\mathrm{p}(X|M,S) = \sum_{i=1}^{k} w_i \, \mathrm{N}(X|M,S_i)$ where S_i is the 2x2 matrix of covariances, $\mathrm{X} = \binom{x_1}{x_2}$ and $M_i = \binom{m_1}{m_2}$. The GMM calculates a "soft" assignment to the modes with the Bayes theorem, which determines the likelihood of X being allocated to one of the k modes for a given value. Parameter optimization methods such as the expectation-maximization (EM) algorithm [1] are commonly utilized in various domains to fit GMMs to two-dimensional data [2-4]. However, automatic modeling of two-dimensional GMMs does not guarantee accurate findings because the EM algorithm is quite sensitive to initial values [5]. As a result, it is advisable to assess the correctness of the derived GMM model through visual means [6]. Therefore, we propose a human-in-the-loop (HIL) approach for modeling two-dimensional gaussian mixtures. Although commercial software approaches exist that provide some range of interactivity for modeling two-dimensional GMMs (e.g., https://www.originlab.com/fileExchange/details.aspx?fid=472: or https://de.mathworks.com/help/stats/tune-gaussian-mixture-models.html), to the authors' knowledge, no fully interactive user interfaces for two-dimensional GMMs have been published so far. Here, we fully integrate the EM optimization of two-dimensional GMM into the interactive adjustment of EM parameters based on visualizations and automatically estimate the number of modes using [7]. To ease the Human-the-loop (HIL) into our interactive tool, we simplify the EM parameters as described in the next section. The third demonstrates that the proposed system “AdaptGauss2D” can improve the automatic state-of-the-art EM modeling of two-dimensional GMMs.

2 System Description

The interactive tool allows the manual modification of all GMM parameters. In this work, the covariance matrix is approximated with the principal component axes (PCA) to ease interactivity for HIL connecting structure of GMM component to parameter axes and angle. Therefore, we propose to compute the principal component axis with a singular value decomposition (SVD) in the first step. Since a two-dimensional space is used, both the diagonal matrix Σ and the unitary matrices U and V^* from the SVD are square matrices. A second step computes ellipsoids based on the two axes for the two-dimensional case. The angle of the axes can be deducted from the axes' position relative to the cartesian coordinate system. The covariance matrix can be computed in a final third step based on the ellipsoid with a rotation matrix. The first step is the SVD of the matrix M resulting from the EM computation with $\mathrm{M}, \mathrm{U}, \Sigma, V^* \epsilon \mathbb{R}^{2\times 2}$. The square root of the singular values p_1 and p_2 denote the length of the principal component axes. The angles can be computed based on the vectors in matrix U and the standard basis vectors. Furthermore, the orientation can be determined by the smaller angle between the main axis and the second vector of the standard basis using $\alpha = \operatorname{acos}\left(\frac{\langle u_1, \mathrm{e}_1\rangle}{\|u_1\|}\right)$. The main axes and the angle define a unique ellipsoid on the cartesian coordinate system, which can be transformed into a symmetric positive definite matrix. A rotation matrix R needs to be defined based on the priorly computed angle and axes with

$$\mathrm{R} = \begin{pmatrix} cos\left(\frac{alpha\cdot\pi}{180}\right) & -sin\left(\frac{alpha\cdot\pi}{180}\right) \\ sin\left(\frac{alpha\cdot\pi}{180}\right) & cos\left(\frac{alpha\cdot\pi}{180}\right) \end{pmatrix} \tag{1}$$

The rotation matrix is applied to transform the matrix P defined by the length of the principal component axes $P = \begin{pmatrix} p_1 & 0 \\ 0 & p_2 \end{pmatrix}$. The symmetric positive definite matrix C



Feature 1: SS

Feature 2: FS

Gaussian No.: 1 of 3: 1

Mean of SS: 5,487

Main Axis 1: 0,5022

Weight: 1: 0,294

Mean of FS: 5,6942

Main Axis 2: 0,4253

Angle: 284,2

Angle of Main Axis 1: 284.2

☐ EM Add/Rem. Modes

EM Steps: 10

RMSD%: 0.4233

Fig. 1. Top Screenshot of the interface of the AdaptGauss2D tool. Bottom: Three-dimensional visualization of the model fitted with the interactive tool in color versus the density estimation of data in black of the flow-cytometry dataset that was interactively modeled after the usage of EM. The third dimensions indicate the density.

can be deployed as the covariance matrix for the original problem with $C = R \cdot P^2 \cdot R^T$. The following actions can be done by the user: 1. Select the Gaussian that you intend to edit. Add new Gaussians or remove the currently selected one. 2. Modify the parameters of the selected Gaussian. The selected Gaussian is shown in blue. 3. Use the knob slider as an alternative way to set the ellipsoids angle of the currently selected gaussian. 4. Use the interactive three-dimensional plots to understand the two-dimensional data and/or model. 5. Use one of the four two-dimensional visualizations to understand the data or model better. 6. Use the upper buttons to switch between a view of the empirical density estimation and the model, switch on or off the data's scatter, or compare the original classification (if given) of the data versus the model’s classification using the Bayes Theorem. Here, the maximum of the posterior distribution is used as a hard classification. 7. Switch between the four different two-dimensional plots (upper button row) and switch on or off the ellipsoid’s axis and outline the models’ components, the data scatter or choose between the original classification (if given) and the model’s classification. 8. Execute the EM algorithm with the desired number of steps

(EM Steps), allow the EM to add/remove modes or not, undo or redo any change made by the algorithm or by hand, preserve the currently selected Gaussian weight, and norm the others or norm all the Gaussian weights. 9. Load a classification to compare with the model's computed classification, save the current setting, or close AdaptGauss2D.

3 Evaluation and Application

In the video https://www.youtube.com/watch?v=MV7DVEWys_c the dataset EngyTime is used, which is described in [8]. The identification of cluster structures combined with an EM algorithm yields a root mean square deviation (RMSD) of 20%. A manual fitting of the initialized results from the automatized adaptation reduced RMSD to 5%. Comparing both solutions to the ground truth shows an improvement of accuracy from 0.921 to 0.965. Fig. 1 presents a flowcytometry sample file of blood with N=296.755 measured event. In Flow Cytometry, each cell rapidly passes through a laser beam one by one. Two light scatter and several surface parameters can be measured for each event. Fig.1 (top) presents the forward scatter FS versus side light scatter SS in which three modes are visible in the shiny interactive app. However, the EM algorithm is unable to fit the Gaussians to the data, as Fig. 1 (top) shows. Here, the identification of cluster structures combined with an EM algorithm yields an RMSD of 0.4233%. A manual fitting by a HIL of the initialized results from the automatized adaptation reduced the RMSD to 0.0605%. The result of the interactive modeling is presented in Fig. 1 bottom as a three-dimensional plot for which the density of the model (colors) and data (back) is shown in the third dimension.

References

1. Baggenstoss PM. Statistical modeling using gaussian mixtures and hmms with matlab. Naval Undersea Warfare Center, Newport RI. 2002.
2. Yoshida E, Kimura Y, Kitamura K, Murayama H. Calibration procedure for a DOI detector of high resolution PET through a Gaussian mixture model. IEEE Transactions on Nuclear Science. 2004;51(5):2543-9.
3. Yu J. Bearing performance degradation assessment using locality preserving projections and Gaussian mixture models. Mechanical Systems and Signal Processing. 2011;25(7):2573-88.
4. Wang et al. Efficient volume exploration using the gaussian mixture model. IEEE Transactions on Visualization and Computer Graphics. 2011;17(11):1560-73.
5. Yang M-S, Lai C-Y, Lin C-Y. A robust EM clustering algorithm for Gaussian mixture models. Pattern Recognition. 2012;45(11):3950-61.
6. Ultsch et al. Identification of Molecular Fingerprints in Human Heat Pain Thresholds by Use of an Interactive Mixture Model R Toolbox (AdaptGauss). International journal of molecular sciences. 2015;16(10):25897-911. doi: 10.3390/ijms161025897.
7. Thrun MC, Stier Q. Fundamental Clustering Algorithms Suite SoftwareX. 2021;13(C):100642. doi: 10.1016/j.softx.2020.100642.
8. Thrun MC, Ultsch A. Clustering Benchmark Datasets Exploiting the Fundamental Clustering Problems. Data in Brief. 2020;30(C):105501. doi: 10.1016/j.dib.2020.105501.