GRay: Ray Casting for Visualization and Interactive Data Exploration of Gaussian Mixture Models – Additional Material

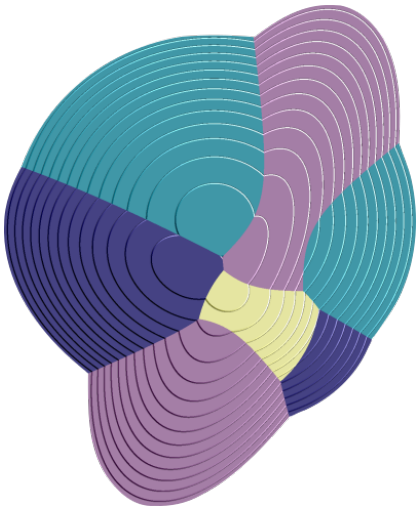
Kai Lawonn, Monique Meuschke, Pepe Eulzer, Matthias Mitterreiter, Joachim Giesen, Tobias Günther

1 Questionnaire

In the following, we describe the questionnaire, including the explanations given to the participants, the images shown, and the questions themselves. The questionnaire began with the collection of general information about the participants (fields of expertise, years of experience, age, gender). Afterwards, an introduction to the tool was given. The questionnaire itself consisted of 17 questions to be answered on a five point Likert scale from 1-strongly disagree to 5-strongly agree. The questions were grouped into several sections.

Maximum Intensity Projection (MIP)

Here, four Gaussian distributions (Gdistr) in the Gaussian Mixture Model (GMM) are visualized in different colors. The maximum value at the camera position along the view direction is determined: values are compared against every Gdistr and the highest value is used for rendering. This yields a Voronoi-like visualization of the GMM that subdivides space into one region for each Gdistr such that the Gdistr has the highest probability in this region among all components. The corresponding Gdistr is shown together with the region. Within each region, we indicate the direction of the corresponding mean vector by the visual cue of stairs. Going upstairs will lead to the mean vector.



Info: We consider this visualization as a first impression of the GMM, it shows a Voronoi-like structure showing which regions have the highest probability to be assigned to a certain Gdistr.

Q1 In MIP, the stairs help to identify the location of the mean, i.e., the center of each Gaussian distribution (Gdistr).

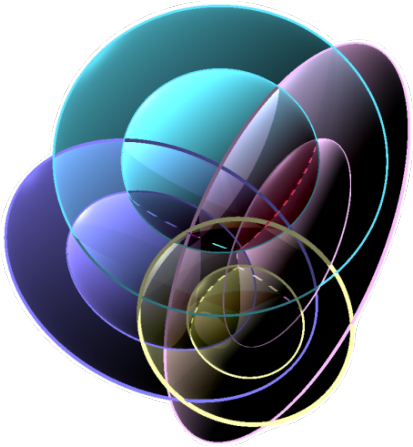
Q2 The MIP visualization helps to understand where the probability to be assigned to a certain Gdistr is the highest, i.e., which regions belong to which Gdistr.

Hull Intersection

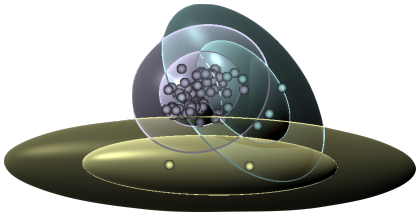
For a 3D impression, we provide a hull visualization. We cast a ray from the camera and identify the intersections for every hull (based on isovalues) and every Gdistr. The 3D impression is supported by transparent hulls and a shading such that nested structures can be identified.

- *K. Lawonn, P. Eulzer, M. Mitterreiter, J. Giesen are with Friedrich Schiller University of Jena. E-mail: first name.last name@uni-jena.de.*
- *M. Meuschke is with Otto von Guericke University of Magdeburg. E-mail: meuschke@isg.cs.uni-magdeburg.de*
- *T. Günther is with Friedrich-Alexander-Universität Erlangen-Nürnberg. E-mail: tobias.guenther@fau.de*

Manuscript received 31 Mar. 2022; accepted 16 Jul. 2022. Date of Publication 26 Sep. 2022; date of current version 26 Sep. 2022. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: 10.1109/TVCG.2022.3209374



Hull rendering with data points assigned to the Gdistr.



Info: This visualization is intended to provide a 3D impression of the Gdistr and data point assignment.

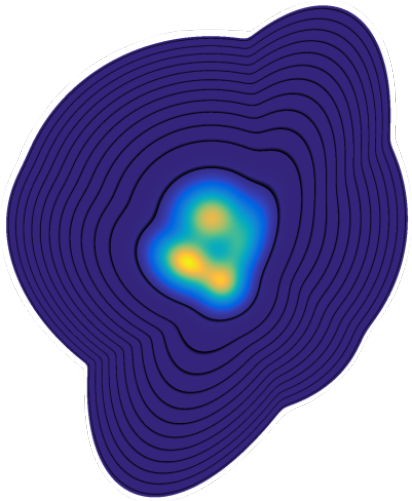
Q3 The hulls help to identify the location of the mean, i.e., the center of each Gaussian distribution (Gdistr).

Q4 The hull visualization helps to understand where the probability to be assigned to a certain Gdistr is the highest, i.e., which regions belong to which Gdistr.

Q5 The dashed lines (intersections of the innermost hulls) support the understanding of the spatial impression of Gdistr.

Direct Volume Rendering

Here, four Gaussian distributions (Gdistr) in the Gaussian Mixture Model (GMM) are visualized. To visualize GMMs, we need to display accumulated values along the ray. This visualization helps to identify modes, i.e., local maxima of the mixture distribution. Note that the modes are not necessarily aligned with the mixture components. Furthermore, we add isolines (varying width) to identify the direction of increasing values.



Info: We consider this visualization for the specific task to explore the GMM and to search for new modes that are emerged by the Gdistr. This visualization is intended for GMM exploration and to search for new modes that emerge from the Gdistr.

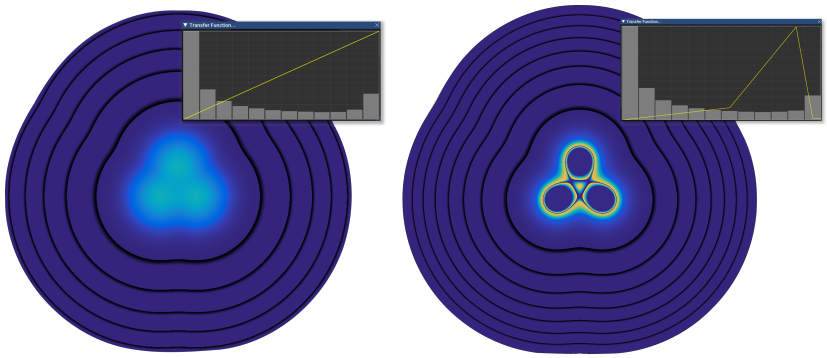
Info 2: In the 2D case the summation of n Gaussian distributions lead to a function with at most (!) n local maximum points. In higher dimensions, it may happen that more (!) than n maximum points occur. The occurrence of maximum points is also referred to as a mode. So a mode is a region of local high values, which may be identified in a cumulative direct volume rendering (DVR).

Q6 The DVR helps to identify if new modes, i.e., regions of high values emerge (if the GMM consists of four Gdistr and five regions of locally high values are visible then a new mode emerge).

Q7 The DVR visualization helps me to understand where the probability is the highest (yellow color).

Q8 The isolines help to identify in which direction the probability increases.

Three Gdistr are shown with a standard transfer function to visualize the GMM on the left. After changing the transfer function (right), a new mode emerges (in the middle a new region with high values occurs).

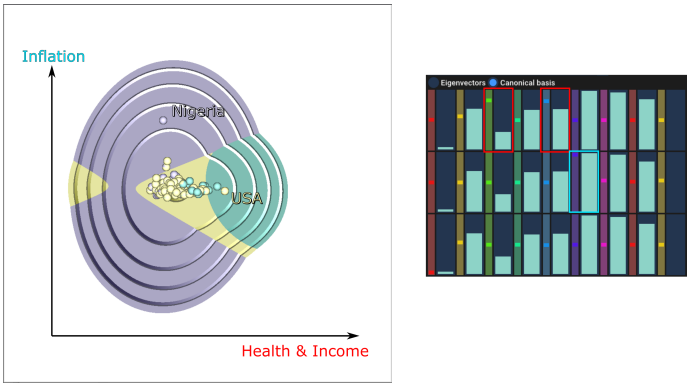


Q9 Changing the transfer function allows to identify new emergent modes.

Basis Exploration

Dataset countries of the world. The x-axis (red) is set to 1 as the first attribute (child mortality), the y-axis (cyan) is set to 1 as the fifth attribute (income). We can immediately see that Haiti has a high child mortality and a low income whereas the situation is swapped in Qatar.

Another example: The x-axis (red) is set to equal parts to the attributes health and income, the y-axis (cyan) is set to the attribute inflation. We can immediately see that the USA have a high income and a high number of healthy citizens and a low inflation, whereas the situation is swapped in Nigeria.

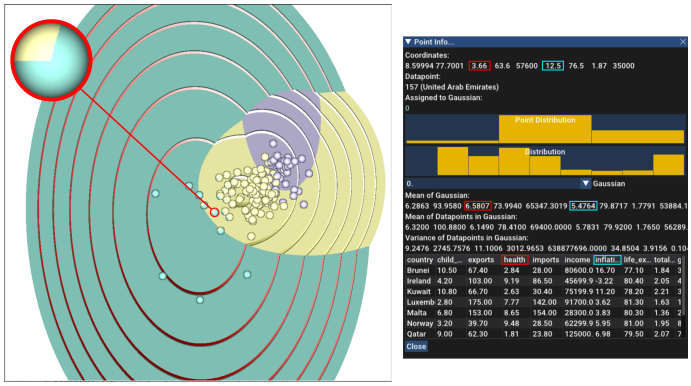


Q10 The basis vector customization allows to set the attribute for the x- and y-axis to gain insights into the data.

Q11 With this, the data can easily be explored.

Point Visualization and Analysis

This dataset shows countries of the world. The highlighted data point (United Arab Emirates (UAE)) belongs to the first Gdistr/cluster (richest countries in the world) displayed in teal. As shown in the distribution with a probability over 25% it belongs to the second cluster of countries displayed in yellow. Analyzing the data, we can see that UAE has a health value (red) of 3.66, which is lower than the average of this cluster (6.6 or 6.1), furthermore the inflation (cyan) is much higher (12.5) compared to the average (5.5 or 5.8) in this cluster. This explains why it has a probability of about 75% of belonging to the first cluster and about 25% of belonging to the second one.

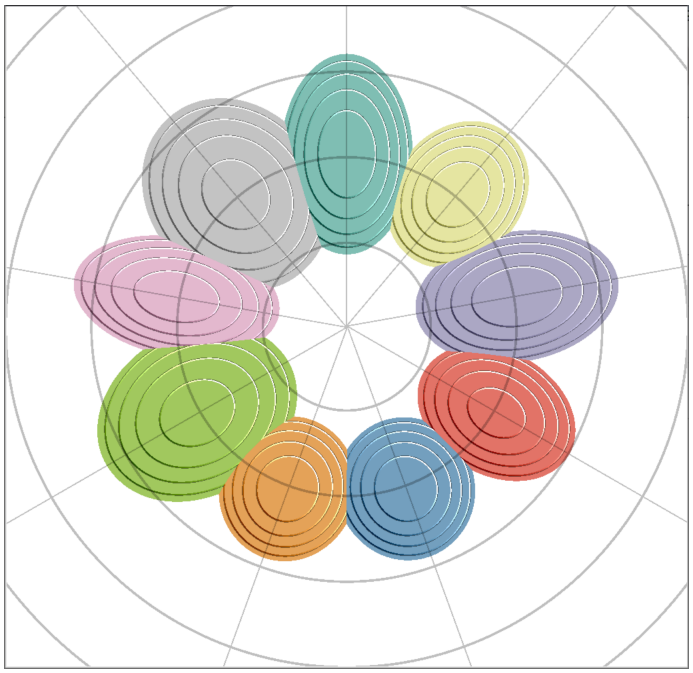


Q12 Using the data point visualization, I can identify whether a data point has a relatively high probability or whether it also belongs to another Gdistr/cluster.

Q13 Using the information box, I can analyze and understand why a data point has a higher probability to be part of another cluster.

Circle Plot

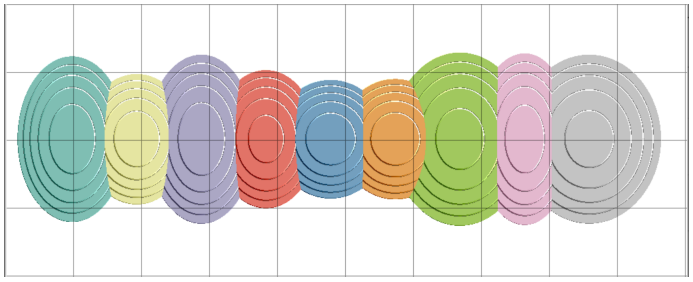
Overview of Gdistr: Gdistr are placed on a circle (axes are aligned with the two eigenvectors with highest eigenvalues) to compare their extents.



Q14 The circle parametrization gives an overview to compare the extents of the Gdistr.

Line Plot

Overview of Gdistr: Gdistr are placed on a line (axes are aligned with the two eigenvectors with highest eigenvalues) to compare their extents.



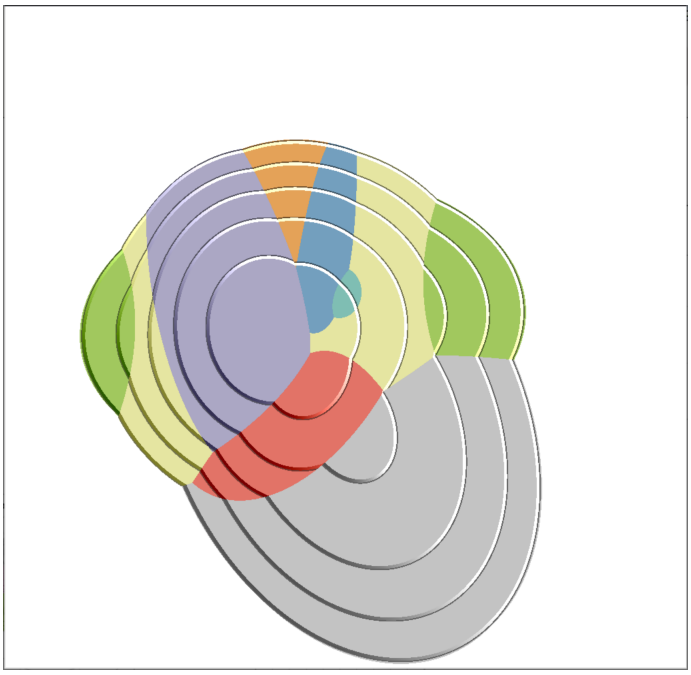
Q15 The line parametrization gives an overview to compare the extents of the Gdistr.

Showing both the circle plot and the line plot again, the following single choice question was asked:

S1 Which plot do you prefer? (Circle, Line, Both)

Principal Component Plot

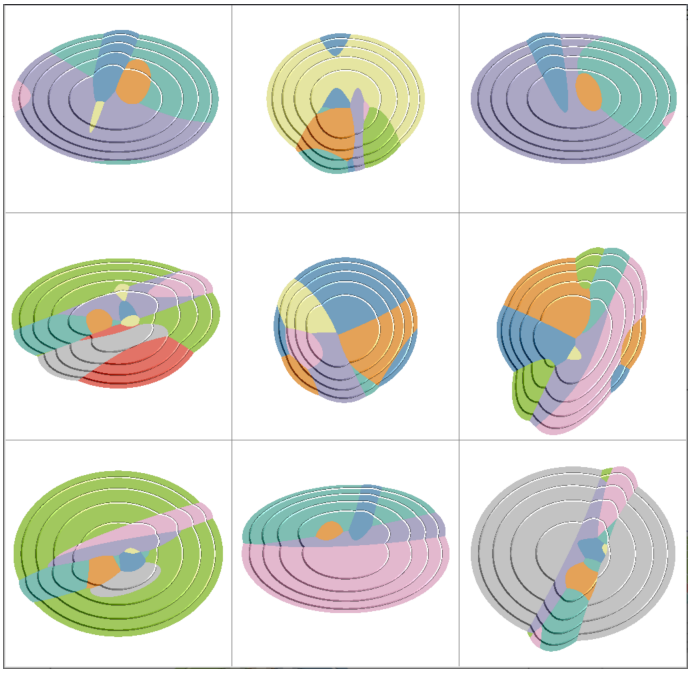
Overview of Gdistr: Gdistr are projected on a 3D subspace with PCA to give an overview.



Q16 The PCA parametrization gives an overview of the GMM.

Small Multiples

Overview of Gdistr: The mean vector of every Gdistr is placed in the middle of a subplot. The axes are aligned with the two eigenvectors with highest eigenvalues.



Q17 The small multiples give an overview of the GMM.

Showing both the principal components plot and the small multiples again, the following single choice question was asked:

S2 Which plot do you prefer? (Principal component plot, Small multiples, Both)

2 Findings in an Additional Data Set

CIFAR-10 In the following, we describe with an additional use case, how our tool can be used to benefit in a scenario of deep learning. Here, we have $N = 14$ Gaussians, $k = 246$ dimensions, and the data lives on a $m \approx 212$ dimensional subspace, as the 14 Gdistrs capture 90% of the information for 214, 208, 215, 212, 210, 215, 214, 211, 212, 213, 212, 211, 214 and 211 eigenvalues, respectively.

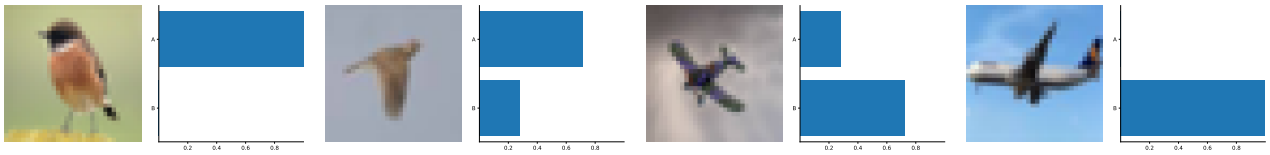


Fig. 1. Four query images from the CIFAR-10 data set together with normalized log probabilities for the respective topics. 1. pure *bird* topic, 2. mixture of *bird*, and *airplane* topic, 3. mixture of *airplane* and *bird* topic, 4. pure *airplane* topic.

Siamese networks [1] are a family of deep unsupervised representation learning methods. Here, we use the SimSiam framework [2] for representation learning on images. The representations are given as feature vectors in $\mathbb{R}^k$, where the embedding dimension k is the most important hyper-parameter. On the feature vectors, we then train a k-variate GMM using the EM algorithm. The GMM is a probabilistic model $p(x,z)$ on the features x and on the latent class variable z that can take the values $1,\ldots,N$. It holds that $p(x,z=i) \sim \phi_i \mathcal{N}(x,\mu,\Sigma)$. We can use the model for two types of inference queries. First, by sorting training or test feature vectors $\hat{x}$ by their density values $p(x|z=i)$, we define a soft cluster or topic on these vectors. Second, for a given feature vector x, we can compute a cluster or topic assignment by

$$\text{argmax}_{i=1,\ldots,N}\, p(z=i|x).$$

We study both types of queries on the CIFAR-10 data set that contains 60 000 labeled RGB images in ten classes with 6 000 images each. The ten classes of the CIFAR-10 data set can be recovered although no information about these classes, not even their number, has been used while learning the feature vectors and while training the GMM. Increasing the number of components can lead to topic splits, that is, one topic is split into two or more related topics. For example, a topic that refers to horses is split into a white horses, horses with riders, and horse heads. Hence, topic splits can be a valuable property of GMM topic models that needs to be tracked over the choice of the number of model components. Ideally, one wants to predict which topics are going to split when the number of components is increased.

In Fig. 1, we show examples of topic assignments for query images that are first transformed into their respective SimSiam feature vectors. For validating the GMM, we are especially interested in query images where the topic assignment is not clear cut. Such query images are typically the hard cases that can impose constraints on the topic model. The two hard cases shown in Figure 1 would likely be also classified correctly by most humans though these are indeed hard instances.

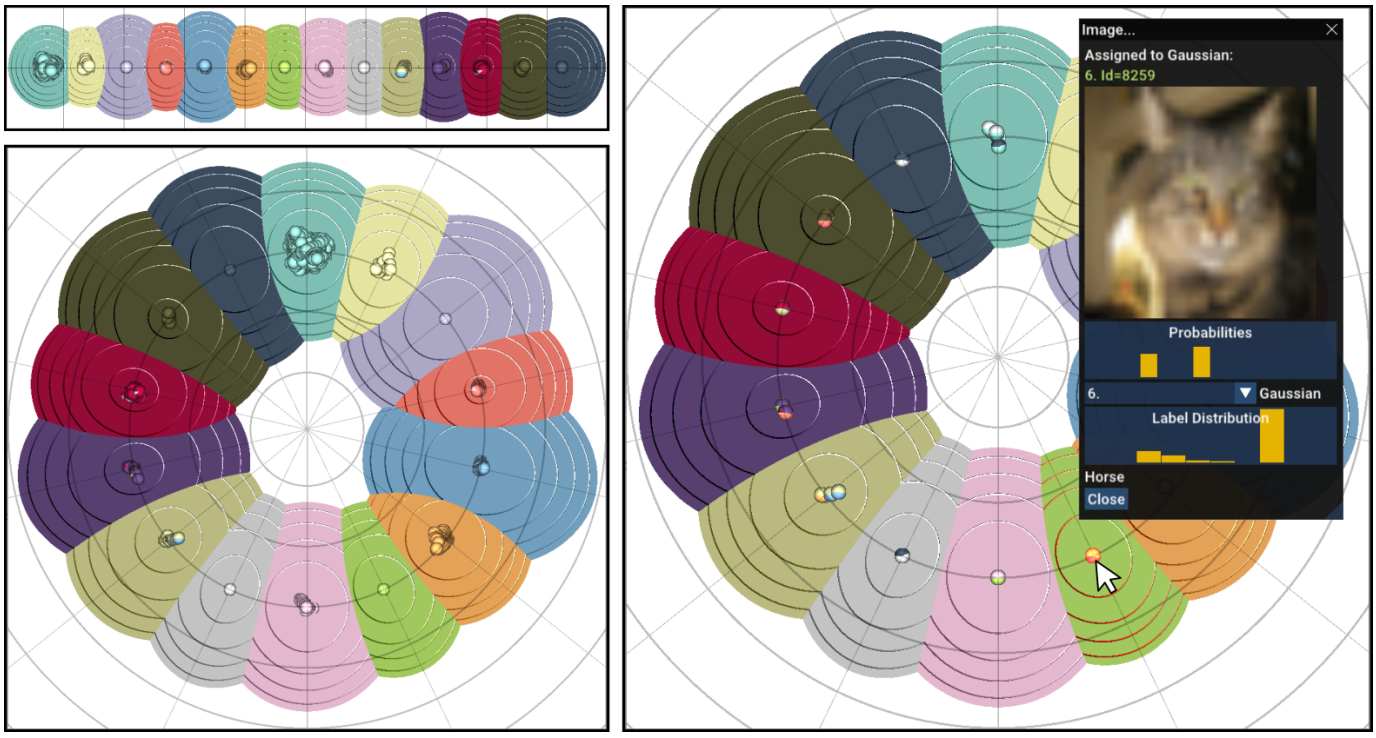
However, exploring topic boundaries, as we did for generating the examples in Figure 1, is a laborious manual task that requires manual coding. This becomes much easier by our interactive tool for exploring GMMs. First, we used the line and the circle plot to gain an overview of the data, see Fig. 2(a) (left). Afterwards, we filtered out the data points that are less than 60% certain to be associated to a Gdistr. Here, we identified a data point and selected it, see Fig. 2(a) (right). We found out that this data point (id=8259) belongs to the sixth Gaussian, which is also shown here. Further, we can see that this data point is an image of a cat and that the sixth Gaussian contains mostly images of horses, which is also stated in the menu under the histogram. We can also explore from the histogram that this selected data point has a 56% likelihood to be in Topic 6 (sixth Gaussian) and a 43% to be in Topic 3. Selecting the third Gaussian, we see that this cluster actually contains mostly images of cats.

In another experiment, we obtained feature vectors of dimension 128 and determined a GMM of three and four Gdistrs. Here, we observed that cats and dogs belong to a single Gdistr and that another Gaussian contains means of transportation, i.e., airplanes, automobiles, ships, and trucks. We also observed that when changing the GMM from three to four Gdistrs, one topic is split into two. The case $N = 3$ includes a topic that contains birds, cats, dogs, and frogs. Increasing N to four, this topic splits and new topic emerges that contains mostly frogs. There still is a topic contains birds, dogs, and cats, see Fig. 2(b).

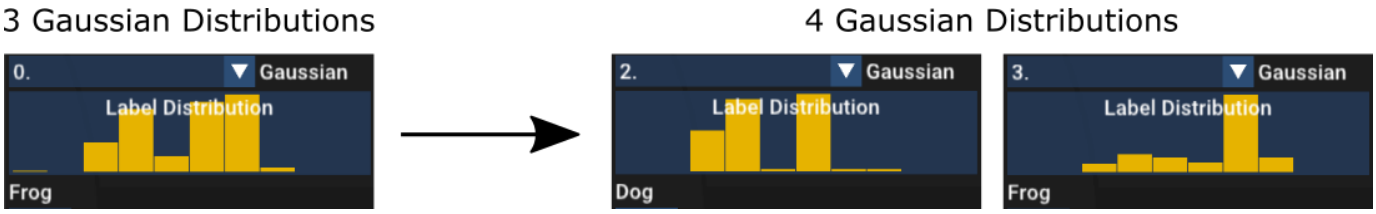
References

[1] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature Verification Using a Siamese Time Delay Neural Network. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 737–744, 1993.

[2] X. Chen and K. He. Exploring simple siamese representation learning. *CoRR*, abs/2011.10566, 2020.



(a)



(b)

Fig. 2. (a) We explored the CIFAR-10 data by first using the line and the circle plot. Afterwards, we filtered the data points to only have points that are less than 60% certain to be associated to a Gdistr. One data point is the image of a cat, but is assigned to a horses topic. (b) Changing the number of Gdistrs in the GMM leads to a split of a Gdistr in two Gdistrs.