MODEL CENTROIDS FOR THE SIMPLIFICATION OF KERNEL DENSITY ESTIMATORS

Olivier Schwander[⋆]*, Frank Nielsen*[⋆†]

[⋆] École Polytechnique, Palaiseau, France
[†] Sony Computer Science Laboratories Inc., Tokyo, Japan

ABSTRACT

Gaussian mixture models are a widespread tool for modeling various and complex probability density functions. They can be estimated using Expectation–Maximization or Kernel Density Estimation. Expectation–Maximization leads to compact models but may be expensive to compute whereas Kernel Density Estimation yields to large models which are cheap to build. In this paper we present new methods to get high-quality models that are both compact and fast to compute. This is accomplished with clustering methods and centroids computation. The quality of the resulting mixtures is evaluated in terms of log-likelihood and Kullback-Leibler divergence using examples from a bioinformatics application.

***Index Terms*—** Kernel Density Estimation, simplification, Expectation–Maximization, k-means, Fisher-Rao centroid

1. INTRODUCTION

Statistical methods are nowadays commonplace in modern signal processing. There are basically two major approaches for modeling experimental data by probability distributions: we may either (1) consider a semi-parametric modeling by a finite mixture model learnt from the Expectation Maximization (EM) procedure, or alternatively (2) choose a non-parametric modeling using a kernel density estimator (KDE).

On the one hand mixture modeling requires to fix or learn the number of components but provides a useful compact representation of data. On the other hand, KDE finely describes the underlying empirical distribution at the expense of the dense model size. In this paper, we present a novel statistical modeling method that simplifies efficiently a KDE model with respect to an underlying distance between Gaussian kernels. We consider Fisher-Rao metric or Kullback-Leibler divergence. Since the underlying Fisher-Rao geometry of Gaussian is hyperbolic without closed-form equation for the centroids, we rather adopt a close approximation that bears the name of hyperbolic model centroid, and show its use in single-step clustering method. We report on our experiments that shows that the KDE simplification paradigm is a competitive approach over the classical EM, both in terms of processing time and quality.

2. MODELING DISTRIBUTIONS

Mixtures of Gaussian distributions are a widespread tool for modeling complex data in a lot of various domains, from image processing to medical data analysis through speech recognition. This success is due to the capacity of Gaussian Mixture Models (GMM) to estimate the probability function (pdf) of complex random variables. For a mixture f of n components, the probability density function takes the form:

$$f(x) = \sum_{i=1}^{n} \omega_i \, g(x;\, \mu_i, \sigma_i^2) \tag{1}$$

where ω_i denotes the weight of component i ($\sum \omega_i = 1$). Each component $g(x;\, \mu_i, \sigma_i^2)$ is a normal distribution with the pdf:

$$g(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}$$

Such a mixture can be built using the celebrated Expectation-Maximization algorithm (EM) which iteratively estimates the parameters that maximize the likelihood.

Kernel Density Estimation (KDE), also known as the Parzen windows method [1], estimates a probability

density function known by N samples using a sum of N kernel functions (usually Gaussian kernels). Each Gaussian in centered at a sample point and has its smoothness controlled by a fixed variance-covariance matrix, called the bandwidth. The density estimate is then a fixed-variance Gaussian mixture model, with N components. See [2] for an extensive survey of the various bandwidth selection methods.

The main drawbacks of the EM algorithm are the risk to converge to a local optimum and the number of iterations needed to find this optimum. While it may be costly, this time is only spent during the learning step. On the other hand, learning a KDE is nearly free but evaluating the associated pdf is costly since we need to loop over each Gaussian of the mixture. Given the typical size of a dataset (a 120×120 image leads to 14400 components), the mixture can be unsuitable for time critical applications. Since mixtures with a low number of components have proved their capacity to model complex data, it would be useful to build such a mixture avoiding the costly learning step of EM.

3. SIMPLIFYING KERNEL DENSITY ESTIMATE

Given a KDE, the most straightforward way to reduce the number of components is to clusterize them, using a k-means-like algorithm. Goldberger and Roweis [3] and Garcia *et al.* [4] already proposed such a method, using the Kullback-Leibler divergence which is well suited for comparing probability density functions. Moreover, the divergence formula and the centroid formula are known in closed form. It allows efficient computation of the two steps of a k-means, repartition and centroids update. Even without closed-form formula, it is sufficient to know an algorithmic way to compute centroids to build a k-means algorithm.

Although the results are quite good (see Figure 1), the problem of the number of iterations remains. Our proposed solution is to only performs the initialization and one iteration of the k-means.

4. METRIC AND NON-METRIC GAUSSIAN CENTROIDS

4.1. Divergence-based centroids

The Kullback-Leibler divergence (KLD) measures the relative entropy between two distributions. For Gaussian distribution, this yields to:

$$\begin{aligned}\mathrm{KLD}(f_p, f_q) = &\frac{1}{2}\log\left(\frac{\det\Sigma_p}{\det\Sigma_q}\right)\\ &+\frac{1}{2}\mathrm{tr}(\Sigma_q^{-1}\Sigma_p)\\ &+\frac{1}{2}(\mu_q-\mu_p)^T\Sigma_q^{-1}(\mu_q-\mu_p)-\frac{d}{2}\end{aligned} \tag{3}$$

where d is the dimension of the space.

Notice that this function is not symmetrical, we need to define three kinds of centroids:

- right-sided one: $\arg\min_c \sum_i \omega_i KLD(c, x_i)$
- left-sided one: $\arg\min_c \sum_i \omega_i KLD(x_i, c)$
- symmetrized one: $\arg\min_c \sum_i \omega_i \mathrm{SKL}(x_i, c)$

with $\mathrm{SKL}(p,q) = \frac{1}{2}\left(KLD(p,q) + KLD(q,p)\right)$

Closed-form formulas are known both for divergence and for centroid but this does not mean that the computation is cheap due to complex operations involved (determinant, matrix inversion, etc).

4.2. Fisher-Rao and model centroids

Given the hyperbolic geometry of the Gaussian distribution, a closed-form formula of the Fisher-Rao distance can be expressed, using the Poincaré hyperbolic distance in the Poincaré upper half-plane:

$$\begin{aligned}&\mathrm{FRD}(f_p, f_q) =\\ &\sqrt{2}\ln\frac{|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_q}{\sqrt{2}},\sigma_q)|+|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_p}{\sqrt{2}},\sigma_p)|}{|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_p}{\sqrt{2}},\sigma_p)|-|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_p}{\sqrt{2}},\sigma_p)|}\end{aligned} \tag{4}$$

In order to perform the k-means iterations using Fisher-Rao distance, we need to define centroids on the hyperbolic space. Model centroids, introduced by Galperin [5], are a way to define centroids in the three

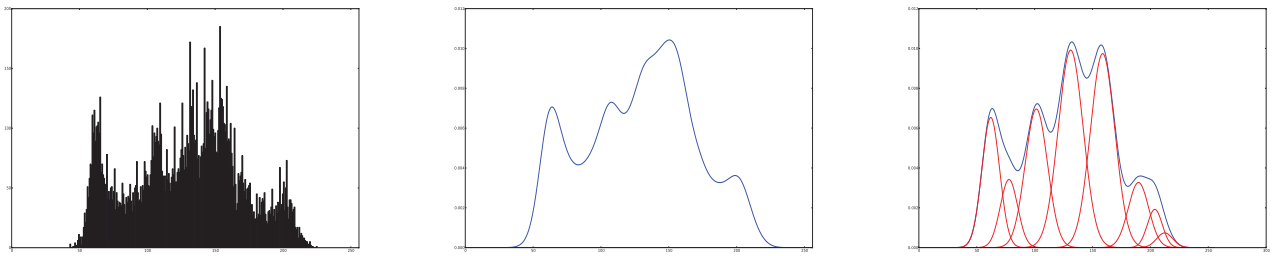

Fig. 1. Original histogram, raw KDE (14400 components) and simplified mixture (8 components)

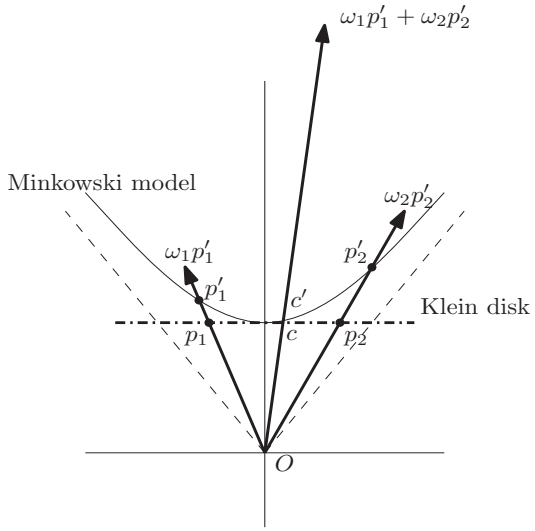

Fig. 2. Computation of the centroid c given the system $(\omega_1, p_1), (\omega_2, p_2)$

kinds of constant curvature spaces (Euclidean, hyperbolic or spherical). For a d dimensional curved space, it starts with finding a $k+1$ dimensional model in the Euclidean space. For a 2D hyperbolic space, it will be the Minkowski model, that is the upper sheet of the hyperboloid $-x^2 - y^2 + z^2 = 1$.

First, each point p (with coordinates (x_p, y_p)) lying on the Klein disk is embedded in the Minkowski model:

$$x_{p'} = \frac{x_p}{1 - x_p^2 + y_p^2} \qquad y_{p'} = \frac{y_p}{1 - x_p^2 + y_p^2} \qquad z_{p'} = \frac{1}{1 - x_p^2 + y_p^2} \tag{5}$$

Next the center of mass of the points is computed

$$c'' = \sum \omega_i p'_i \tag{6}$$

This point needs to be normalized to lie on the Minkowski model, so we look for the intersection between the vector Oc'' and the hyperboloid:

$$c' = c''/(-x_{c''}^2 - z_{c''}^2 + z_{c''}^2) \tag{7}$$

From this point in the Minkowski, we can get a point in the Klein disk:

$$x_c = \frac{x_{c'}}{z_{c'}} \qquad y_c = \frac{y_{c'}}{z_{c'}} \tag{8}$$

Although this scheme gives the centroid of points located on the Klein disk, it is not sufficient since parameters of the Gaussian distribution are in the Poincaré upper half-plane. Thus we need to convert points from one model to another, using the Poincaré disk as an intermediate step. For a point (a, b) on the half-plane, let $z = a + ib$, the mapping with the Poincaré disk is:

$$z' = \frac{z - i}{z + i} \qquad z = \frac{i(z' + 1)}{1 - z'} \tag{9}$$

And for a point k on the Poincaré disk, the mapping with a point k on the Klein disk is:

$$p = \frac{1 - \sqrt{1 - \langle k, k \rangle}}{\langle k, k \rangle} \qquad k = \frac{2}{1 + \langle p, p \rangle} p \tag{10}$$

5. EXPERIMENTS

We study here the quality, in terms of log-likelihood, and the computation time of the proposed methods compared to a baseline Expectation-Maximization algorithm. The source distribution is the intensity histogram of the famous Lena image. For the Kullback-Leibler divergence, we report only results for right-sided centroids since, as reported in [4], it performs better than the two other flavors and has the same computation cost. The third method is the Model centroid, both with a full k-means and with only one iteration.

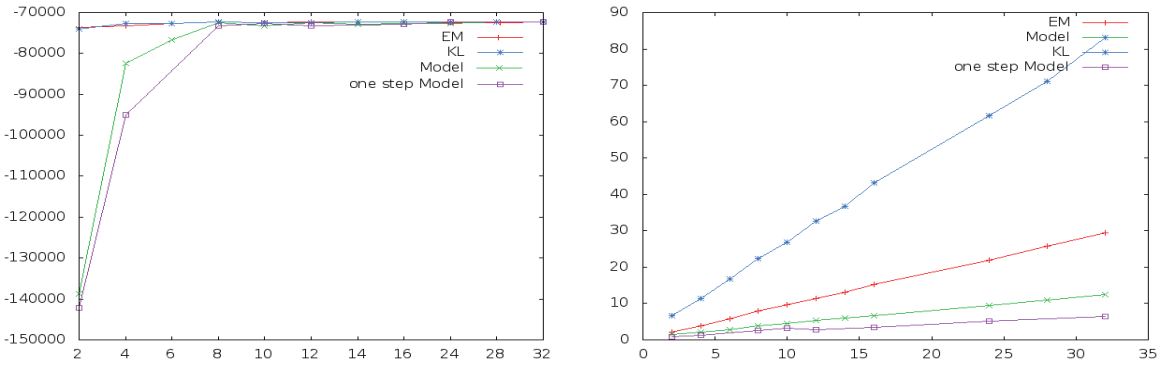

Fig. 3. Log-likelihood of the simplified models and computation time

The left hand side of Figure 3 shows the evolution of the log-likelihood as a function of the number of components k. First, we see that all the algorithms perform nearly the same and converge very quickly to a maximum value (the KL curve is merged with the EM one).

The right part of Figure 3 describes the running time (in seconds) as a function of k. Despite the fact that the quality of mixtures is nearly identical, the costs are very different. Kullback-Leibler divergence is very slow (even in closed-form, the formulas are quite complex). While achieving same log-likelihood, model centroid is the fastest method, significantly faster than EM.

While being slower to converge when k increases, the one step model clustering performs still well and is roughly two times faster than a complete k-means. The initialization is random: we do not use k-means++ here since its cost during initialization cancels the benefit of performing only one step.

On the dataset from [6] we study the Kullback-Leibler divergence between the original KDE and the simplified mixtures: mixtures from model centroid simplification and from Kullback-Leibler simplification give comparable results and are 4 to 10 times better than Expectation-Maximization.

6. CONCLUSION

We presented here a novel modeling paradigm which is fast and accurate. From a precise but difficult to use model, the kernel density estimate, we are able to build new models which achieve the same approximation quality while being faster to compute and easier to use and which are able to outperform Expectation-Maximization. Moreover, we show that Model centroids are good candidates for approximating Fisher centroids.

Source code, additional materials and experiments are available online [1].

7. REFERENCES

[1] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[2] S.J. Sheather and M.C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 53, no. 3, pp. 683–690, 1991.

[3] J. Goldberger and S. Roweis, "Hierarchical clustering of a mixture model," *Advances in Neural Information Processing Systems*, vol. 17, pp. 505–512, 2005.

[4] V. Garcia, F. Nielsen, and R. Nock, "Levels of details for gaussian mixture models," *Computer Vision–ACCV 2009*, pp. 514–525, 2010.

[5] G.A. Galperin, "A concept of the mass center of a system of material points in the constant curvature spaces," *Communications in Mathematical Physics*, vol. 154, no. 1, pp. 63–84, 1993.

[6] J. Bernauer, X. Huang, A.Y.L. Sim, and M. Levitt, "Fully differentiable coarse-grained and all-atom knowledge-based potentials for rna structure evaluation," *RNA*, vol. 17, no. 6, pp. 1066, 2011.

[1] http://www.lix.polytechnique.fr/~schwander/icassp2012