COMIX: JOINT ESTIMATION AND LIGHTSPEED COMPARISON OF MIXTURE MODELS

Olivier Schwander[1,2]*, Stéphane Marchand-Maillet*[2]*, Frank Nielsen*[3,4]

[1] Sorbonne Universités, UPMC Univ Paris 06, LIP6, Paris, France
[2] University of Geneva, CVML/CUI, Carouge, Switzerland
[3] Paris-Saclay University, Ecole Polytechnique, LIX, Palaiseau, France
[4] Sony Computer Science Laboratories, Tokyo, Japan

ABSTRACT

The Kullback-Leibler divergence is a widespread dissimilarity measure between probability density functions, based on the Shannon entropy. Unfortunately, there is no analytic formula available to compute this divergence between mixture models, imposing the use of costly approximation algorithms. In order to reduce the computational burden when a lot of divergence evaluations are needed, we introduce a sub-class of the mixture models where the component parameters are shared between a set of mixtures and the only degree-of-freedom is the vector of weights of each mixture. This sharing allows to design extremely fast versions of existing dissimilarity measures between mixtures. We demonstrate the effectiveness of our approach by evaluating the quality of the ordering produced by our method on a real dataset.

Index Terms— Mixture model, Density estimation, Information geometry, Kullback-Leibler divergence, Exponential family

1. INTRODUCTION AND MOTIVATION

The Kullback-Leibler divergence [1] (KL) is the relative Shannon entropy between two probability density functions:

$$\mathrm{KL}\left(p\|q\right) = \int p(x)\,\log\frac{p(x)}{q(x)}\mathrm{d}x \tag{1}$$

$$= H(p,q) - H(p) \tag{2}$$

where $H(p)$ is the Shannon entropy and $H(p,q) \geq H(p)$ is the cross-entropy. The divergence between p and q measures the amount of information which is lost when q (coming from an estimation algorithm) is used to approximate p (the true distribution).

Since it is not symmetrical, it cannot be a distance but it still bears a subset of the properties of the distance: self similarity ($\mathrm{KL}\left(p\|p\right) = 0$); self identification ($\mathrm{KL}\left(p\|q\right) = 0 \Rightarrow p = q$); positivity ($\mathrm{KL}\left(p\|q\right) > 0$). Moreover it has the property of invariance to reparameterization [2] and is infinitesimally related to the Fisher-Rao-Hotelling distance [3, 4, 5] of information geometry. It is also a Bregman divergence, allowing to apply tools from computational information geometry such as Bregman Voronoi diagrams [6].

For all these reasons, the Kullback-Leibler divergence is a widespread tool in many statistical modeling, signal processing and pattern recognition applications: source separation [7], speech recognition [8], background extraction for object tracking in videos [9].

There is a closed-form formula available when the divergence is computed between two members of the same exponential family (such as between two Gaussian distributions), using the bijection between exponential families and Bregman divergences [10]. But for two mixture models with $k_1, k_2 > 1$ components

$$m(x) = \sum_{i=1}^{k_1} \omega_i p_F(x;\eta_i) \quad m'(x) = \sum_{i=1}^{k_2} \omega'_i p_F(x;\eta'_i) \tag{3}$$

an analytic form does not exist[11] (unlike for other divergences such as Squared Loss[12] and Cauchy-Schwartz[13]). We choose here to focus on mixtures of exponential families p_F: the loss of generality is very weak since a lot of common mixtures are exponential families (including Gaussian) and it allows to build generic algorithms where the family is just a parameter[14].

When a lot of divergence evaluations are needed, the cost of the faster approximation techniques may not

be affordable: the best compromise between speed and accuracy, the variational approximation [15], needs to evaluate the $k_1 \times k_2$ Kullback-Leibler divergences between each pair of components of the two mixtures $\mathrm{KL}\left(p_F(\cdot;\eta_i)\|p_F(\cdot;\eta'_i)\right)$. In order to reduce the computational burden, we introduce a sub-class of mixture models where the parameters of the components are shared between all the mixtures and the only degree-of-freedom is the vector of weights of each mixture (the set of shared components can thus be seen as a dictionary and the weights as the activation of the different atoms of the dictionary). In this case, the previous divergences between pairs of components can thus be precomputed just after the estimation of the mixtures, limiting the computational cost during the evaluations of the variational approximation of the divergence.

This article is organized as follows: after this introduction and a description of previous works on Kullback-Leibler approximation, we define the concept of co-mixtures and then introduce an Expectation-Maximization-based algorithm to build such mixtures; next we introduce fast variants of the variational and Goldberger approximations; finally we experimentally study the quality of our mixture estimation and the speed-up of the approximation on a simulated retrieval application built on top of a real bio-informatics dataset.

2. COMPARING MIXTURE MODELS

A lot of work have been devoted to overcome the lack of an analytic formula for the Kullback-Leibler divergence, we review here the most important in practical applications. One of the major methods relies on Monte-Carlo integration to estimate the divergence. For the random variates $x_1, \ldots, x_n$ drawn from the mixture m, the Monte-Carlo estimator (KLMC) is:

$$\mathrm{KL}_{\mathrm{MC}}\left(m\|m'\right) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{m(x_i)}{m'(x_i)} \tag{4}$$

This formula has the advantage of being consistent but requires a large number of variates to achieve a good precision.

Instead of trying to numerically compute the integral, another approach is to design new functions which are both close to Kullback-Leibler and computable in closed-form. A first example is the Goldberger approximation[8] which comes from a majoration of the Kullback-Leibler divergence:

$$\mathrm{KL}_{\mathrm{Gold}}\left(m\|m'\right) = \arg\min_{\sigma}\mathrm{KL}\left(\omega\|\sigma(\omega')\right) + \sum\omega_i\mathrm{KL}\left(p_F\left(\cdot\|\eta_i\right)\|p_F\left(\cdot\|\eta'_{\sigma(i)}\right)\right) \tag{5}$$

In this approximation, a minimization problem is solved over all the possible permutations σ of the components in order to cope with all the possible orderings of the components of the two mixtures.

A better approximation in the state-of-the-art is the variational approximation[15]:

$$\mathrm{KL}_{\mathrm{var}}\left(m\|m'\right) = \sum_i \omega_i \log\frac{\sum_j \omega_j e^{-\mathrm{KL}(p_F(\cdot\|\eta_i)\|p_F(\cdot\|\eta_j))}}{\sum_j \omega'_j e^{-\mathrm{KL}\left(p_F(\cdot\|\eta_i)\|p_F\left(\cdot\|\eta'_j\right)\right)}} \tag{6}$$

This approximation achieves the lowest variance compared to a KLMC with a very larger number of points (more than 1 million) but is not positive.

3. DEFINITION AND ESTIMATION

Definition 1. A co-mixture of exponential families (a *comix*) with K components is a set of S statistical mixture models of the form:

$$\begin{cases} m_1(x;\omega_i^{(1)}\ldots\omega_K^{(1)}) = \sum_{i=1}^K \omega_i^{(1)} p_F(x;\eta_i) \\ m_2(x;\omega_i^{(2)}\ldots\omega_K^{(2)}) = \sum_{i=1}^K \omega_i^{(2)} p_F(x;\eta_i) \\ \ldots \\ m_S(x;\omega_i^{(S)}\ldots\omega_K^{(S)}) = \sum_{i=1}^K \omega_i^{(S)} p_F(x;\eta_i) \end{cases} \tag{7}$$

p_F is the exponential family with log-normalizer F and $\eta_1 \ldots \eta_K$ are the parameters of the components and are shared between all the individual mixtures of the co-mixture; the S vectors $\omega_1^{(l)}\ldots\omega_K^{(l)}$ are the vectors of weights (thus positive) with the property $\sum_{i=1}^K \omega_i^{(l)} = 1$ for any l. The parameter S depends on the number of sets of points which are modeled jointly.

In order to estimate the parameters of the co-mixture (the vectors of parameters components and the matrix of weights), we adapt the Bregman Soft Clustering algorithm[10], which is a variant of EM for exponential families. The version for co-mixtures, called co-Expectation-Maximization (co-EM) searches for a local minimum of the average of the log-likelihoods of the S individual mixtures on the associated input set of points $\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(S)}$:

$$L(\mathcal{X}^{(1)}, \ldots, \mathcal{X}^{(S)}) = \frac{1}{S}\sum_{l=1}^{S} l^{(l)}(\mathcal{X}^{(l)}) \tag{8}$$

This is done with an iterative EM-like algorithm classically divided into two main steps:

Expectation step We compute S responsibility matrices $p^{(1)}, \ldots, p^{(S)}$:

$$p^{(l)}(i,j) = \frac{\omega_j^{(l)} p_F(x_i^{(l)}, \eta_j)}{m(x_i^{(l)}|\omega^{(l)}, \eta)} \tag{9}$$

Maximization step The maximization step is itself divided in two steps: first, S partial estimates (not shared) are computed then all these partial estimates are combined into the new estimate of the shared parameters.

Weights and partial estimates for the *l*-th dataset are computed using the observations for $\mathcal{X}^{(l)}$ and the *l*-th responsibility matrix:

$$\eta_j^{(l)} = \sum_i \frac{p^{(l)}(i,j)}{\sum_u p^{(l)}(u,j)} t(x_i^{(l)}) \tag{10}$$

$$\omega_j^{(l)} = \sum_i \frac{p^{(l)}(i,j)}{\sum_u p^{(l)}(u,j)} \tag{11}$$

For the *j*-th component, the new estimate of η_j is computed as a Bregman barycenter of all the $\eta_j^{(1)}, \ldots, \eta_j^{(S)}$, giving the same weight to all the sets of points:

$$\eta_j = \frac{1}{S} \sum_{l=1}^{S} \eta_j^{(l)} \tag{12}$$

Notice that we work with the expectation parameters η of the exponential family.

4. FAST DIVERGENCES

The dissimilarity measures between mixtures presented previously can be rephrased to take into account the shared parameters. As the terms computing KL between identical components vanish, the Goldberger approximation becomes the following simpler formula: $\mathrm{KL}_{\mathrm{Gold}}(m\|m') = \mathrm{KL}(\omega\|\omega')$. The minimization problem also vanishes since we do not have any more to cope with the different orderings of the components.

For the variational approximation, the terms involving the cross dissimilarity between all pairs of components can be precomputed since the parameters are known in advance: $D_{ij} = \mathrm{KL}(p_F(\cdot\|\eta_i)\|p_F(\cdot\|\eta_j))$.

5. EXPERIMENTS

From the practitioner point of view, the absolute values of a dissimilarity measure are not really important: relative values are a lot more valuable in most cases. Moreover, the largest values of a divergence are often of low interest since the useful information is usually concentrated in the closest points instead of the farthest ones. In order to demonstrate both the quality of the mixture estimation using comix and the quality of the divergence computations, we thus study in which measure the ordering of the points is the same between a reference baseline and comix-based methods. This evaluation is made by simulating a retrieval application and measuring the mean average precision (mAP) over all the possible queries (by successively taking each mixture as the query and looking at the retrieved mixtures in a short list of size 10). The experiments presented here are made on a bio-informatics 1D dataset[16, 17]: it consists of 211 sets of points (so the parameter S will be 211) with 3000 to 5000 observations in each set. We estimate a density for each set of points with a Gaussian Mixture Model. The chosen baseline is made of an estimation step with classical EM (with 8 components, as set by an expert) and a comparison step made with KLMC (with 1 million variates, which is often considered enough to have a very small variance on the estimation).

Fig. 1 (left) describes the evolution of the precision of the KLMC estimation with respect to the sample size and compares it with the precision of the variational KL approximation. A target objective to validate the quality of the proposed joint estimation method is to be as good as the variational approximation between EM mixtures. Fig. 1 (right) studies the precision of our fast versions of the KL approximations with respect to the number of components. Even if some value is considered a good number of components for individual mixtures, it may need a lot more components to cope with the variations of all the set of points during the joint estimation. We see here that 4 components are sufficient to achieve the same precision rate for variational KL on comix and variational KL on traditional mixtures. It is even more interesting to see that with 32 components we get precision values which are only attained with the lot more costly Monte-Carlo method. The Goldberger approximation has a similar behavior but outperforms variational KL for all number of components. Our interpretation is that the joint estimation of a co-mixture

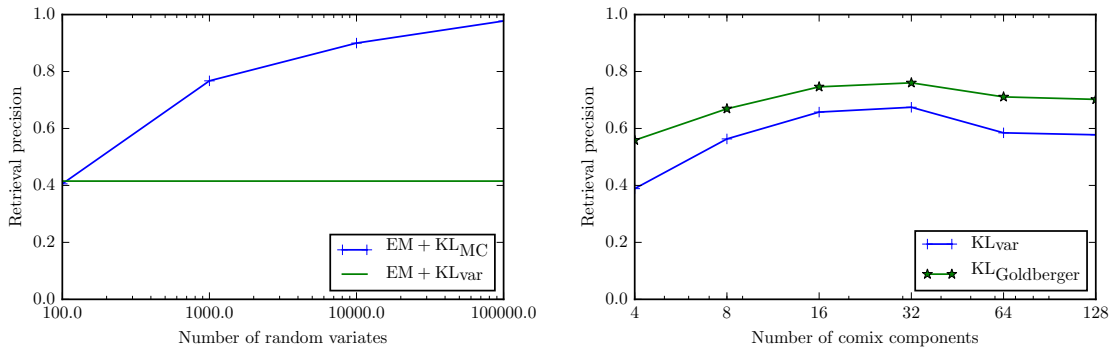

Fig. 1. *Left*: mAP of KLMC between EM mixtures wrt the sample size and result from variational KL. *Right*: mAP wrt the number of components of variational Kullback-Leibler and Goldberger between co-EM mixtures.

k	co-EM	Speed-up between co-EM and EM8	KL_{var} on comix	Speed-up between KL_{var} on comix and KL_{var} on EM8	Speed-up between KL_{var} on comix and KL_{MC100} on EM8	Goldberger on comix
4	51s	$\times$1.5	0.00020s	$\times$180	$\times$ 20	0.00015s
8	99s	$\times$0.77	0.00044s	$\times$84	$\times$ 5.8	0.00030s
16	48s	$\times$1.6	0.0012s	$\times$28	$\times$ 1.6	0.00059s
32	150s	$\times$0.49	0.0040s	$\times$9.1	$\times$ 0.41	0.0012s
64	450s	$\times$0.17	0.014s	$\times$2.5	$\times$ 0.10	0.0024s
128	600s	$\times$0.12	0.046s	$\times$0.80	$\times$0.026	0.0049s

Table 1. Absolute times for computation on comix and speed-up when compared to the times of the equivalent computation on individual mixtures. Times for co-EM are compared with the total time for all the individual EM.

is able to capture intrinsic similarities between set of points which are missed out when the sets are looked at independently by EM.

Table 1 shows the computation times for co-EM and for the fast divergence computations and displays the speed-up compared to the estimation of the 211 mixtures with EM and the speed-up compared to slow version of the approximations. We first see that the computation cost of the estimation of a comix with co-EM on all the sets of points has the same order of magnitude than the estimation of all the individual mixtures with EM, meaning that there is no significant loss of time during the estimation step. Comparison between fast variational KL and variational KL between traditional mixtures shows a big speed-up for a number of components between 4 and 64, meaning that the method would cope with a larger number of components if needed. The Goldberger approximation has a similar cost than variational Kullback-Leibler, without precomputation step, making it very interesting due to its better performances. All computation were made on an Intel i5-4440 CPU.

6. CONCLUSION

We presented a new approach in the field of the approximations of the Kullback-Leibler divergence between mixtures: instead of designing estimation techniques or introducing different but related formulas which are easy to compute, we work on the nature itself of the mixture models, by imposing the shared parameters among all the considered mixtures. This sharing allows to build extremely fast versions of classical Kullback-Leibler approximations: experiments on a retrieval task show that the combination of co-mixtures and fast approximations are not only fast but also meaningful. Nevertheless, we are able to outperform the retrieval precision attained by methods working on individual mixtures: this allows a large set of exploration paths to understand and exploit how the notion of co-mixture can improve performances in fields where mixtures are commonly used.

Acknowledgement Part of this work has been supported by the Swiss National Science Foundation via Project MAAYA (Grant number 144238).

7. REFERENCES

[1] Solomon Kullback, *Information theory and statistics*, Courier Corporation, 1997.

[2] Shun ichi Amari and Hiroshi Nagaoka, *Methods of information geometry*, vol. 191, American Mathematical Soc., 2007.

[3] R. A. Fisher, "Two new properties of mathematical likelihood," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 144, no. 852, pp. 285–307, Mar. 1934.

[4] C. Radhakrishna Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bulletin of the Calcutta Mathematical Society*, vol. 37, no. 3, pp. 81–91, 1945.

[5] H Hotelling, "Spaces of statistical parameters," *Bull. Amer. Math. Soc.*, vol. 36, pp. 191, 1930.

[6] F. Nielsen, J. D. Boissonnat, and R. Nock, "Bregman Voronoi diagrams: Properties, algorithms and applications," *Institut National de Recherche en Informatique et en Automatique (INRIA Sophia Antipolis), Research Report*, vol. 6154, 2007.

[7] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[8] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 487–493.

[9] Chris Stauffer and W Eric L Grimson, "Adaptive background mixture models for real-time tracking," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.* IEEE, 1999, vol. 2.

[10] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *The Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[11] S Watanabe, K Yamazaki, and M Aoyagi, "Kullback information of normal mixture is not an analytic function," *Technical Report of IEIE in Japanese*, , no. 2004-0, pp. 41–46, 2004.

[12] Meizhu Liu, Baba C Vemuri, S-I Amari, and Frank Nielsen, "Total Bregman divergence and its applications to shape retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3463–3468.

[13] Frank Nielsen, "Closed-form information-theoretic divergences for statistical mixtures," in *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012, pp. 1723–1726.

[14] Frank Nielsen and Vincent Garcia, "Statistical exponential families: A digest with flash cards," *CoRR*, vol. 09114863, 2009.

[15] J.R. Hershey and P.A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, Apr. 2007, vol. 4, pp. IV–317–IV–320.

[16] Julie Bernauer, Xuhui Huang, Adelene YL Sim, and Michael Levitt, "Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation," *RNA*, vol. 17, no. 6, pp. 1066–1075, 2011.

[17] Adelene Y. L. Sim, Olivier Schwander, Michael Levitt, and Julie Bernauer, "Evaluating mixture models for building RNA knowledge-based potentials," *Journal of Bioinformatics and Computational Biology*, vol. 10, no. 02, Apr. 2012.