FAST AGGREGATION OF STUDENT MIXTURE MODELS

Ali El Attar, Antoine Pigeau and Marc Gelgon

Nantes university, LINA (UMR CNRS 6241), Polytech'Nantes
rue C.Pauc, La Chantrerie, 44306 Nantes cedex 3, France
phone: + (33) 2 40 68 32 57, fax: + (33) 2 40 68 32 32
email: firstname.lastname@univ-nantes.fr

ABSTRACT

Studies on Mixtures of Student (t-)distributions have demonstrated their ability to conduct clustering tasks with valuable robustness to outliers, compared to their Gaussian mixture counterparts. Concurrently, distributed clustering has motivated much interest in methods for building a partition by consensus of multiple partitions. This paper addresses the latter need by aggregating mixtures of Student distributions. It involves minimizing iteratively an approximate KL divergence between mixtures, which themselves approximate each Student component as a finite Gaussian mixture.

1. INTRODUCTION

Probabilistic mixture models form a mainstream approach to unsupervised clustering, with a wealth of variants, pertaining to the form of the model, optimality criteria and estimation schemes. While Gaussian mixtures are, by far, the most popular, they are known to lack statistical robustness, i.e. estimation of parameters is severely affected by only a modest proportion of outliers. Improvement may be obtained by specifying a clustering-oriented optimality criterion, which encourages well-separated classes, rather than density modelling (max. likelihood like). Another direction for improvement resides in the form of the mixture components. In particular, mixture of t-distributions (i.e. Student distributions) have demonstrated their effectiveness to face this robustness issue, thanks to their heavier tail that can model a larger amount of outliers, compared to Gaussian densities. In fact, as we shall detail, a Student density may be viewed as a infinite linear combination of Gaussians with constant mean and various variances. Smaller variances model the "meaningful" data, while large-variance components account for outliers, if needed. The degree of freedom controls how the amount of outliers to be accepted in the model. The price to pay is more intricate (often, lack of close form) and possibly unstable estimation schemes, especially when the degree of freedom is unknown. This robustness has attracted much interesting into adapting existing important Gaussian-based procedures to their Student-based counterparts [1, 7].

Aggregation of class models is a classical topic, both supervised (ensemble methods, boosting) and unsupervised. Yet, growing interest comes from the transposition of existing statistical learning and recognition tasks onto distributed computing systems (cluster, P2P, sensor networks). The unsupervised case is our focus here, i.e. search for a consensus from an ensemble of data partitions. This issue has been addressed e.g. by probabilistic approaches in the case of Gaussian mixtures [3, 5] and voting techniques [2].

A combined model could simply be obtained by a weighted sum of mixtures, yet this would generally result in an unnecessarily high number of components, with a view to capturing the underlying probability density. The scope of the paper is a new scheme for estimating, from such a possibly over-complex Student mixture, a mixture that is more parsimonious, i.e. where each class is represented by a single component. Parsimony is particularly important if such mixture combinations follow one after another.

Section 2 presents a baseline algorithm, which was proposed to carry out mixture reduction in the case of Gaussian mixtures. Briefly stated, it operates like a k-means technique on mixture components, trying to minimize Kullback-Leibler divergence between the linear combination of incoming mixtures and the reduced mixtures. For the robustness reasons discussed above, it is better suited to density estimation than clustering. Thus, efficient consensus techniques in an ensemble of model-based remains an important and open issue. Section 2 identifies the difficulties in generalizing the baseline mixture reduction algorithm to Student mixtures. There are two main difficulties, that reside in the two steps of the iterative algorithm. Section 3 recalls the expression of a Student density as a Gaussian mixture, and hence of a Student mixtures as a constraint mixture of mixtures. We then put forward solutions of the two abovementionned difficulties. Section 4 provides experimental results and Section 5 supplies concluding remarks.

2. BASELINE MIXTURE REDUCTION ALGORITHM

Let the problem be formulated as transforming a mixture model f into another g with less components, while minimizing the KL divergence involved by the simplification process. This section recalls how this was solved in [5], as we shall progress from this baseline. A key feature of their solution is that only model parameters are accessed to group components, i.e. neither access to data nor sampling are required. Thus, it is very cost effective in terms of computation. The central mechanism in the component grouping technique consists in approximating the KL-divergence between two mixtures f and g as follows :

$$d(f,g) = \sum_{i=1}^{K} \pi_i \min_{j=1}^{M} KL(f_i||g_j) \quad (1)$$

where K and M are respectively the number of components of f and g, π_i is the mixing proportion of component i.

The search for optimal g is composed in two alternating steps. The first one associates components of f to the current components of g (binary assignments), minimizing eq.(1). In other words, it amounts to determining the best mapping m from $\{1 \dots K\}$ to $\{1 \dots M\}$ such that criterion (1)

© EURASIP, 2009

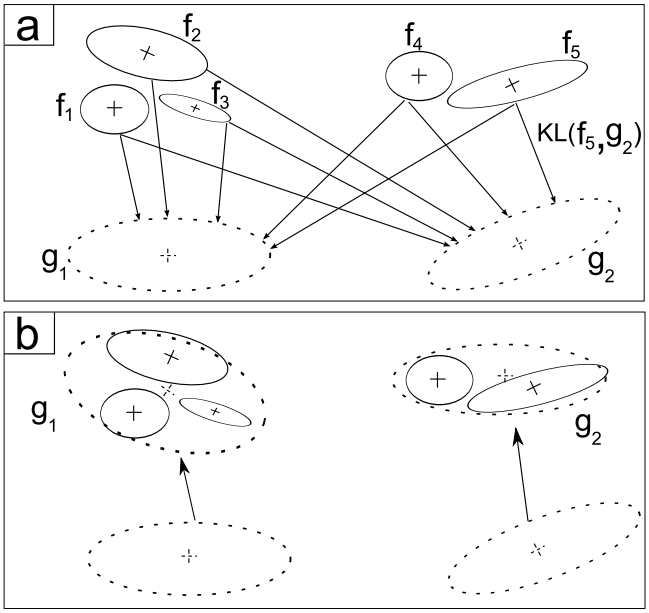


Figure 1: Reduction of a mixture model following [5] : dotted and solid ellipses represent respectively the g and f models. (a) shows the first step where the divergence between components of g and f are computed (see arrows). (b) presents the parameter update of g based on the mapping m, minimizing criterion (1).

is minimized :

$$\begin{aligned} d(f,g) &= \arg\min_{m} \quad d(f,g,m) \\ &= \arg\min_{m} \sum_{i=1}^{K} \pi_i KL(f_i||g_{m(i)}) \end{aligned} \tag{2}$$

The above approximation is practically interesting because there exists the following closed-form approximation of the divergence between Gaussians :

$$\begin{aligned} KL(f_i||g_j) = \quad & \tfrac{1}{2}[log\tfrac{|\Sigma_{g_j}|}{|\Sigma_{f_i}|} + Tr|\Sigma_{g_j}^{-1}\Sigma_{f_i}| \\ & -\gamma + (\mu_{f_i} - \mu_{g_j})^t \Sigma_{g_j}^{-1}(\mu_{f_i} - \mu_{g_j})] \end{aligned} \tag{3}$$

where γ is the dimension of the feature space.

The second step updates the model parameters of g, again from the sole model parameters of f. These two steps are iterated until the convergence of the criterion defined in equation 1. Figure 1 depicts the clustering algorithm.

Adapting this framework to reduction of a Student mixture raises two problems :

- in step 1 : to our knowledge, there does not exist any closed-form solution for the KL divergence between two Student components (at least when the covariance matrices are not proportional). We thus propose (section 3.2) an approximation of the KL divergence, based on a decomposition of a Student component with a finite sum of Gaussian components;
- in step 2 : we should design a low-cost and statistically optimal scheme for determining a Student component that represents the set of grouped Student components.

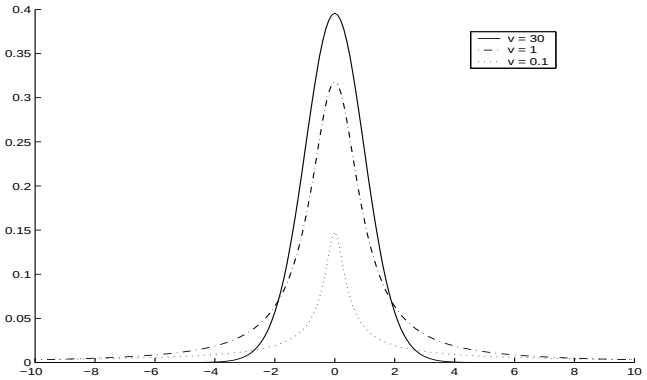


Figure 2: Student density for several values of the degrees of freedom. As $\nu \to \infty$, the distribution corresponds to a Gaussian. For a low degree of freedom, the heavy tails enables more robustness to outliers.

The algorithm 1 summarizes the different steps of our approach.

Algorithm 1 Clustering algorithm for Student components

Require: two Student mixtures f and g, respectively of K and M components ($K > M$). Initial values for means μ_{g_j} are draw randomly and initial covariances for Σ_{g_j} are set to I/p ($1 < p < P$).
1.Approximate each Student component of f and g with P Gaussian components
while $d(f,g)$ is not minimized **do**
2.1.compute the Kullback-Leibler divergence approximation between the components of f and g.
2.2.update the model parameters of g based on the mapping functions m and m'.
end while
Return mixture g, which is the reduction of mixture f

3. REDUCING AN OVERCOMPLEX STUDENT MIXTURE

3.1 Approximation of the Student distribution

A Student distribution S may be expressed as an infinite sum of Gaussian distributions (a Gaussian scale mixture) with identical mean :

$$S(x,\mu,\Sigma,\nu) = \int_0^\infty \mathcal{N}(x,\mu,\Sigma/u)G(u,\nu/2,\nu/2)du, \tag{4}$$

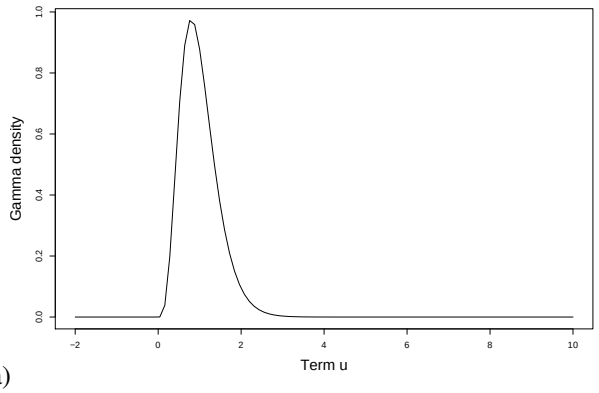
where $\mathcal{N}(x,\mu,\Sigma/u)$ is a Gaussian component with mean μ and covariance Σ, ν is the degrees of freedom and G is the Gamma distribution. Figure 2 presents the curve of a Student distribution as the function of the degrees of freedom. Expression (4) shows that a Student density is an infinite mixture of Gaussians, where weights and covariance are determined by one another.

We propose to approximate each Student component by the following finite mixture of P Gaussians, where $u_1,\ldots,u_P$ are draw from $G(u,\nu/2,\nu/2)$:

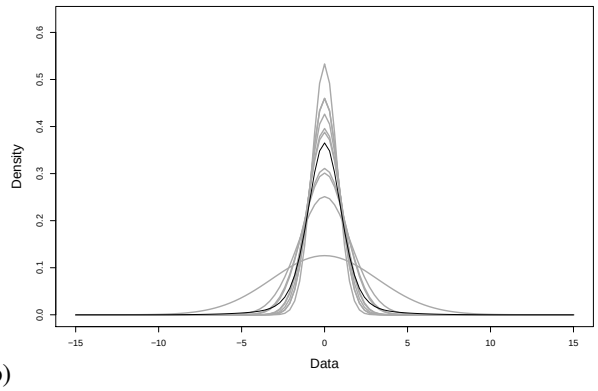
$$S(x) = \frac{1}{P}\sum_{p=1}^{P} \mathcal{N}(x,\mu,\frac{\Sigma}{u_p}) \tag{5}$$

The term P depends on ν, since the higher is ν, the lower P needed for a correct approximation. Each Student component is then approximated with a number of components varying in accordance with its degree of freedom ν.

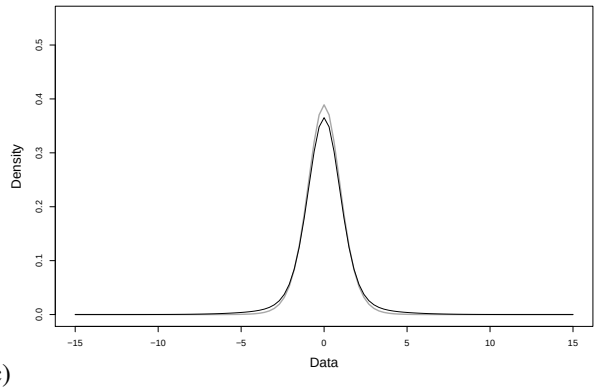
Figure 3 presents an example on our approximation of a Student component with 10 Gaussian components.



(a)



(b)



(c)

Figure 3: (a) presents the Gamma distribution of the term u for $\nu = 10$. Ten values of u are sampled from this distribution. (b) shows the selected Gaussian components (gray lines) and the obtained approximation (black line). (c) compares the approximation (black line) and the real Student density (gray line).

3.2 Kullback-Leibler divergence between two Student approximations

Our Gaussian representation of a Student component is motivated by the opportunity to use an approximation of the Kullback-Leibler divergence between Gaussian mixtures. Although the latter problem has no closed-form expression, it has been given much interest in recent years. Many approaches were compared in [6]. Monte Carlo sampling obviously leads to best accuracy, but at the price of a high calculation complexity. Thus, we rather use it as a benchmark and focus on methods aiming the best trade-off between accuracy in KL approximation and computational cost (i.e. requiring only models parameters). Experiments from [6] conclude that the best approaches are the matched bound and the variational approximations. Because of its lower cost, we resort to the matched bound criterion [4].

This approximation of the Kullback-Leibler divergence between two models f and g is very similar to the previous criterion 2, also based on a mapping function m' minimizing the sum of Kullback-Leibler divergences:

$$KL_{matchBound}(f||g) = \sum_i \pi_i \left(KL(f_i||g_{m'(i)}) + log \frac{\pi_i}{\pi_{m'(i)}} \right). \quad (6)$$

where π_i is the prior probability of a component i.

Approximation of a Kullback-Leibler divergence between two Student components amounts then to computing the Kullback-Leibler divergence between two Gaussian models, both composed of P components and identical means. Figure 4 illustrates our method for computing an approximate Kullback-Leibler divergence between two Student components.

Once the Kullback-Leibler divergence obtained for each approximate Student between f and g, each component of g is assigned to its closest components of f. Parameters of g are then updated in accordance with the mappings m and m'.

3.3 Update of the model parameters

We discussed above how to approximate the KL divergence between mixtures. Since g is unknown, a major point is whether we can associate, to this approximation, a procedure to optimize it. A parameter update scheme was proposed in [5] between Gaussian mixtures. We extend it to cope with Student mixtures. In an iterative fashion, it successively assigns components to groups and updates group representative components. This may be viewed as a k-means like technique operating on mixture components.

Because we approximate each Student component as a finite, constrained Gaussian mixture, the above iterative scheme includes a new inner step to update the parameters of its Gaussian components. Each Gaussian is updated as follows:

$$\mu_{g_j} = \frac{1}{\pi_{g_j}} \sum_{i \in m^{-1}(j)} \pi_{f_i} \mu_{f_i} \quad (7)$$

$$\Sigma_{g_{jp}} = \frac{1}{\pi_{g_j}} \sum_{i \in m^{-1}(j)} \pi_{f_i} \left(\frac{1}{|m'^{-1}(p)|} \sum_{l \in m'^{-1}(p)} (\Sigma_{f_{il}} + (\mu_{f_i} - \mu_{g_j})(\mu_{f_i} - \mu_{g_j})^T) \right) \quad (8)$$

where $\pi_{g_j} = \sum_{i \in m^{-1}(j)} \pi_{f_i}$ and $\Sigma_{g_{jp}}$ is the p^{th} covariance of the component j of g.

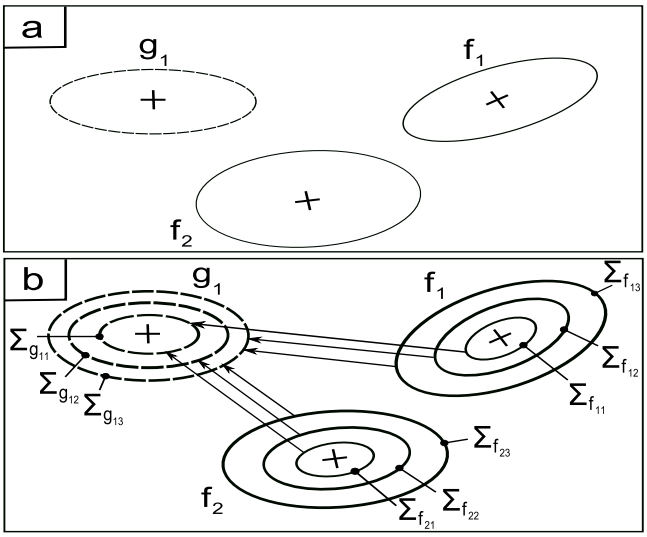


Figure 4: Example of the Kullback-Leibler divergence between two approximate Student components. Solid and dotted lines represent respectively the models f and g. On figure (1) the original Student components of f and g. On figure (2), our proposed approximation of the Student component with $P = 3$ Gaussian components. Optimization of the matched bound criterion amounts to map the components of f and g such that the sum of the Kullback-Leibler divergences is minimized. Here, the arrows show the obtained mapping m'. Note that the mapping is not necessarily surjective.

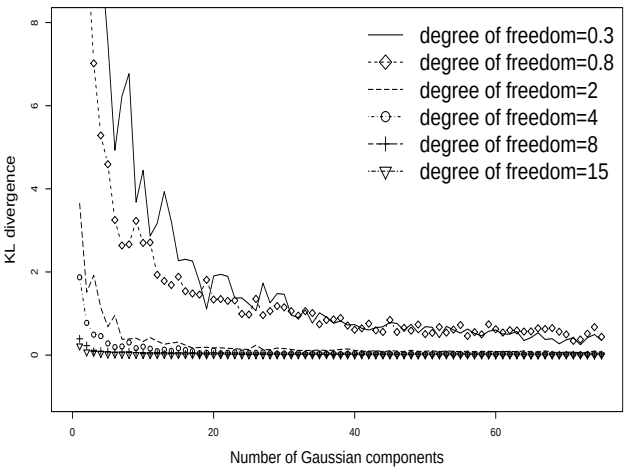


Figure 5: KL divergence between our approximation and the Student distribution according to the number of components and the degrees of freedom ν. As ν increases, the number of components needed to obtain a low KL divergence decreases. This is explained by the fact that the distribution tends to a Gaussian when $\nu \to \infty$.

For a component of g, its single center is the average of the associated centers of f. Each one of its P covariances is the average of the associate covariances of f, based on the mapping m'. For example, the mean and the covariance $\Sigma_{g_{13}}$ of the component g_1 on Figure 4 are respectively updated with the mean of f_1 and f_2 and the covariance $\Sigma_{f_{23}}$, $\Sigma_{f_{11}}$ and $\Sigma_{f_{12}}$.

Note that since m' is not surjective, a component (μ_j, Σ_{jp}) can be associated to no component of f. In this case, the covariance remains unchanged.

4. EXPERIMENTS

To validate our proposal, we first compute a KL divergence between a Student and our approximation for different number of Gaussian components. Then we propose an example of a Student model reduction with our adapted algorithm.

4.1 Approximation of a Student Component

For our first experiment, we sample 5000 data in accordance with a Student distribution and compute the KL divergence based on the Monte Carlo method. For P varying from 1 to 75, we carried out the following steps, 20 times each:

- select the P Gaussian components: randomize P values of term u in accordance with the Gamma distribution;
- compute the KL divergence.

The average KL divergence for the 20 iterations are plotted on Figure 5, for various values of ν.

This experiment confirms that as ν increases, the number of components to obtain a low KL divergence decreases. Indeed, for $\nu \geq 2$, the associated curves tend quickly to 0 giving a good approximation for P varying between 8 and 20 components. For a smaller value of ν, we notice that the divergence converges from 40 Gaussian components.

The chaotic result obtained for the first values on the curves is due to the low number of components used to approximate the Student density. Indeed, the terms u are sampled and each one has a significant weight in these first iterations.

4.2 Reduction of a Student model

We apply here our algorithm to reduce a Student mixture f into a mixture g. Parameters of f are initially set as follows:

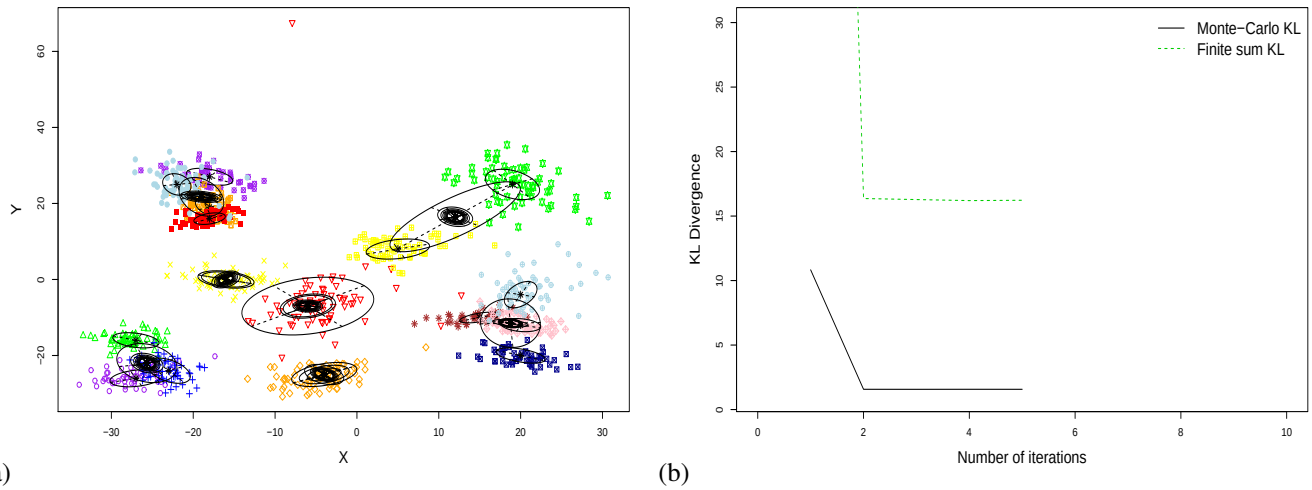
- the mean and the covariance matrix of the initial Student mixture are set manually. To cope with singular covariance, the SVD method is applied on the covariance matrix to factorize them in three matrices $(U\Sigma V^*)$;
- the degree of freedom are drawn uniformly between 0 and 30. Each Student component of f is approximated in accordance with our previous result: if $\nu_i < 3$ then $P_i = 50$ else $P_i = 10$;

Parameters of g are initially set as follows:

- the means are set randomly among the initial means of f;
- the number of Gaussian components P is set with the highest value used to approximate the components of f;
- each covariance of an approximate component are set randomly from the covariance matrices of f.

In our experiment, the model f is composed of 16 Students components (see figure 6 (a)). We present here a reduction solution with a model g composed of 7 components.

The obtained reduction is illustrated in figure 6(a). Overlapped components are well reduced with one component of g (components situated at $(-25,-22)$, $(-18,20)$ and $(20,-15)$), and a similar result is obtained for isolated components (components situated at $(-5,-22)$, $(-7,-10)$ and



(a) (b)

Figure 6: (a) presents the model f and the obtained reduced model g. Single ellipses are the component of f and multiple ellipses are the component of g. The initial data of each component of f are plotted. (b) shows the evolution of KL divergence between f and g during the iterative optimization process. We display both the finite sum approximation, which is more tractable and hence used for optimization, and a Monte-Carlo approximation, which is closer to the true KL.

$(-18, 0)$). We notice an exception with the isolated components $(5, 5)$ and $(20, 20)$, associated to a single component of g. This result, due to the chosen number of components to reduce f (7 components), is nevertheless acceptable considering the proximity of their data sets.

Figure 6 (b) presents the optimization of the Monte Carlo and KL divergences. The optimization is quickly achieved: the first iteration leads to a stable mean for each component and the remaining iterations are due to update of the covariances of each component. The variations on the last iterations are then negligible.

In our opinion, the obtained model g proposes a pertinent reduction of the initial model f. Distinct components are kept and overlapped components are regrouped.

5. CONCLUSION

Student mixtures are powerful tools for robust clustering. This paper proposes a technique for addressing consensus between Student-based partitions, with low-cost, parameter-level. It exploits the view of a Student distribution as an infinite, Gamma-weighted, Gaussian mixture and relies on the iterative optimization of a tractable approximation of a well-founded criterion, KL divergence.

Several improvements are under way. Currently, mixtures to be aggregated are first summed, then the sum is reduced. The reduction process does not yet take into account from which mixture each component originates. Further, a Bayesian treatment will enable a more efficient determination of the suitable number of clusters. The finite sum approximation should make it derivation quite straightforward from [3].

Acknowledgment

This work was partly funded by Region Pays de la Loire (MILES project).

REFERENCES

[1] C. Archambeau and M. Verleysen. Robust Bayesian clustering. *Neural Networks*, 20(1):129–138, 2007.

[2] H. Ayad, O. A. Basir, and M. Kamel. A probabilistic model using information theoretic measures for cluster ensembles. In *Multiple Classifier Systems, 5th International Workshop, MCS 2004, Cagliari, Italy, June 9-11, 2004*, pages 144–153, 2004.

[3] P. Bruneau, M. Gelgon, and F. Picarougne. Parameter-based reduction of Gaussian mixture models with a variational-Bayes approach. In *19th International Conference on Pattern Recognition (ICPR'08)*, Tampa (Fl.), USA, 2008.

[4] J. Goldberger, S. Gordon, and H. Greenspan. An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In *Proceedings of Ninth IEEE International Conference on Computer Vision*, volume 1, pages 487–493, 2003.

[5] J. Goldberger and S. Roweis. Hierarchical clustering of a mixture model. In *Advances in Neural Information Processing Systems 17*, pages 505–512, Cambridge, MA, 2004. MIT Press.

[6] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–317–IV–320, 2007.

[7] M. Svensen and C. Bishop. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005.