

Hierarchical Gaussian Mixture Models

Vincent Garcia, Frank Nielsen, and Richard Nock

ICASSP 2010 – March 14 - 19, 2010 – Dallas, Texas, USA

INTRODUCTION

MIXTURE MODELS

A mixture model f is a powerful framework to estimate probability density functions:

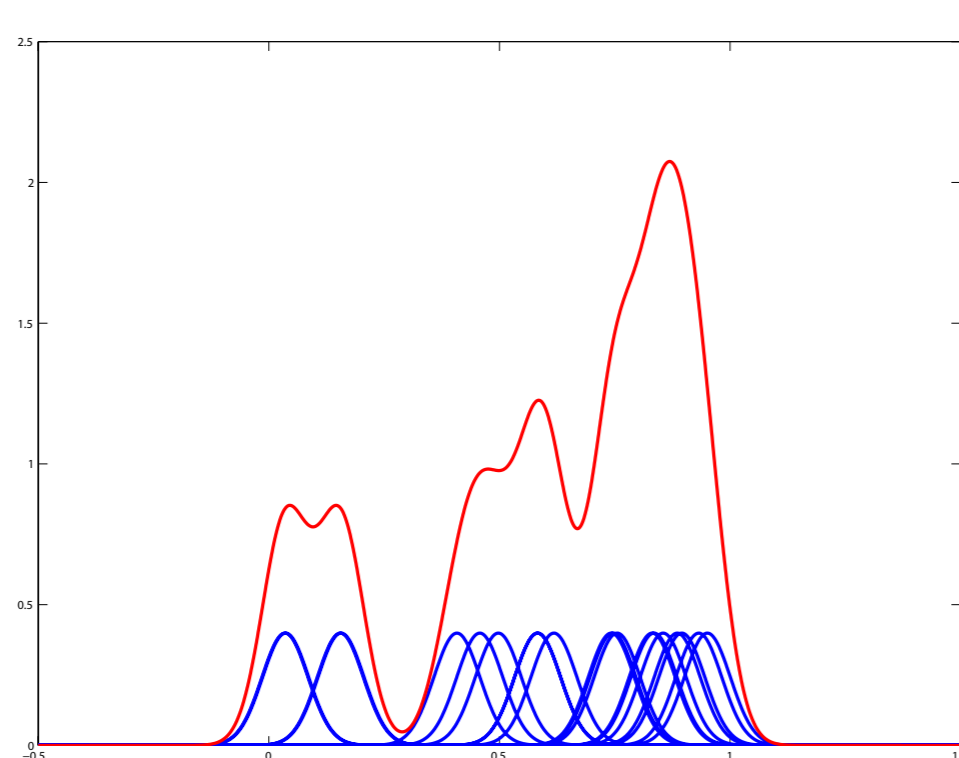
$$f(x) = \sum_{i=1}^n \alpha_i f_i(x) \quad (1)$$

where $\alpha_i \geq 0$ denotes a weight with $\sum_{i=1}^n \alpha_i = 1$. If f is a Gaussian mixture model (GMM),

$$f_i(x) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{(x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i)}{2}\right) \quad (2)$$

with μ_i mean and Σ_i covariance matrix. n denote the number of components.

PROBLEM OF USING MIXTURE MODELS



Density estimation using kernel-based Parzen estimator

Mixture models usually contain a lot of components. The estimation of statistical measures is computationally expensive. We need to reduce the number of components in f . Solutions:

1. Re-learn a simpler mixture model from dataset \Rightarrow **too long and the initial dataset may be not anymore available**
2. Simplify the mixture model $f \Rightarrow$ **most appropriated solution**

Simplifying mixtures to the same number of components allows one to consider Fisher-Rao riemannian geometry too.

MIXTURE MODEL SIMPLIFICATION

Given a mixture model f of n components (see Eq. (1)), the problem of mixture model simplification consists in computing a mixture model g of m components

$$g(x) = \sum_{j=1}^m \beta_j g_j(x)$$

such that g is the *best* approximation of f . The problem is how to simplify f and what is a good value for m ? In this paper, we propose a new algorithm which

1. creates a hierarchical mixture of exponential families denoted h ,
2. provides a progressive representation of f using h ,

MIXTURE OF EXPONENTIAL FAMILIES

KULLBACK-LEIBLER DIVERGENCE AND BREGMAN DIVERGENCE

The fundamental measure between statistical distributions is the Kullback-Leibler divergence (KLD). Given f_i and f_j two distributions, the KLD is a **sided similarity measure** given by

$$D_{\text{KL}}(f_i || f_j) = \int f_i(x) \log \frac{f_i(x)}{f_j(x)} dx \quad (3)$$

In the case of normal distributions

$$D_{\text{KL}}(f_i || f_j) = \frac{1}{2} \log \left(\frac{\det \Sigma_j}{\det \Sigma_i} \right) + \frac{1}{2} \text{tr} \left(\Sigma_j^{-1} \Sigma_i \right) + \frac{1}{2} (\mu_j - \mu_i)^\top \Sigma_j^{-1} (\mu_j - \mu_i) - \frac{d}{2} \quad (4)$$

Normal distributions belong to the class of exponential families. The canonical form is:

$$f_F(x; \Theta) = \exp \{ \langle \Theta, t(x) \rangle - F(\Theta) + k(x) \} \quad (5)$$

The expressions of Θ , $t(x)$, $F(\Theta)$, and $k(x)$ for classical distributions are reported in *Statistical exponential families: A digest with flash cards*. Frank Nielsen, and Vincent Garcia. ArXiv, November 2009.

The KLD between two *density* members of the same exponential family is equal to the Bregman divergence on swapped natural *parameters* and defined for the log normalizer F :

$$D_{\text{KL}}(f_i || f_j) = D_F(\Theta_j || \Theta_i) = F(\Theta_j) - F(\Theta_i) - \langle \Theta_j - \Theta_i, \nabla F(\Theta_i) \rangle \quad (6)$$

Using this formalism, we are able to propose algorithms suitable for the wide class of **mixtures of exponential families**. In particular, this class includes the Gaussian mixture models.

SIDED AND SYMMETRIZED CENTROIDS

Given a set of Gaussians (or a set of member of the same exponential family)

- Right centroid = centroid according to the right-sided Bregman divergence (= left KL centroid).
- Left centroid = centroid according to the left-sided Bregman divergence (=right KL centroid).
- Symmetrized centroid: computed from the right and the left centroids.
- More details in *Sided and Symmetrized Bregman Centroids*, F. Nielsen and R. Nock, *IEEE Transactions on Information Theory*, 2009

BREGMAN HIERARCHICAL CLUSTERING

HIERARCHICAL CLUSTERING

Hierarchical clustering is a method consisting in building a hierarchical clustering of a set of objects (points, etc.). Let P be a set of objects and let P_1, \dots, P_n be a partition of P . Let us consider a distance $D(\cdot, \cdot)$ (called **linkage criterion** and potentially asymmetric) between two subsets.

1. Determine the two closest subsets P_i and P_j relatively to $D(\cdot, \cdot)$
2. Merge P_i and P_j into a single subset
3. If the number of subset is more than one, go to 1.

This algorithm creates a hierarchical structure (dendrogram) containing the merging information.

BREGMAN HIERARCHICAL CLUSTERING ALGORITHM

Let f be a GMM f of size n . If we consider f as a set of n weighted Gaussians (α_i, Θ_i) , we can easily adapt the classical hierarchical clustering to GMM (and to mixtures of exponential families).

1. Determine the two closest subsets of weighted distributions P_i and P_j relatively to $D(\cdot, \cdot)$
2. Merge P_i and P_j into a single subset of weighted distributions
3. If the number of subset is more than one, go to 1.

The linkage criterion $D(\cdot, \cdot)$ is the weighted sided Bregman divergence (right-sided, left-sided, or symmetric). The dendrogram created is called **hierarchical mixture of exponential families**.

PROGRESSIVE REPRESENTATION

From the hierarchical mixture of exponential families h , we can quickly simplify f into a mixture of m components:

1. Extract from h the m subsets of weighted distributions remaining after $n - m$ iterations
2. g_j is the Bregman centroid of the j^{th} extracted subset
3. α_j is the sum of the weights of the j^{th} subset

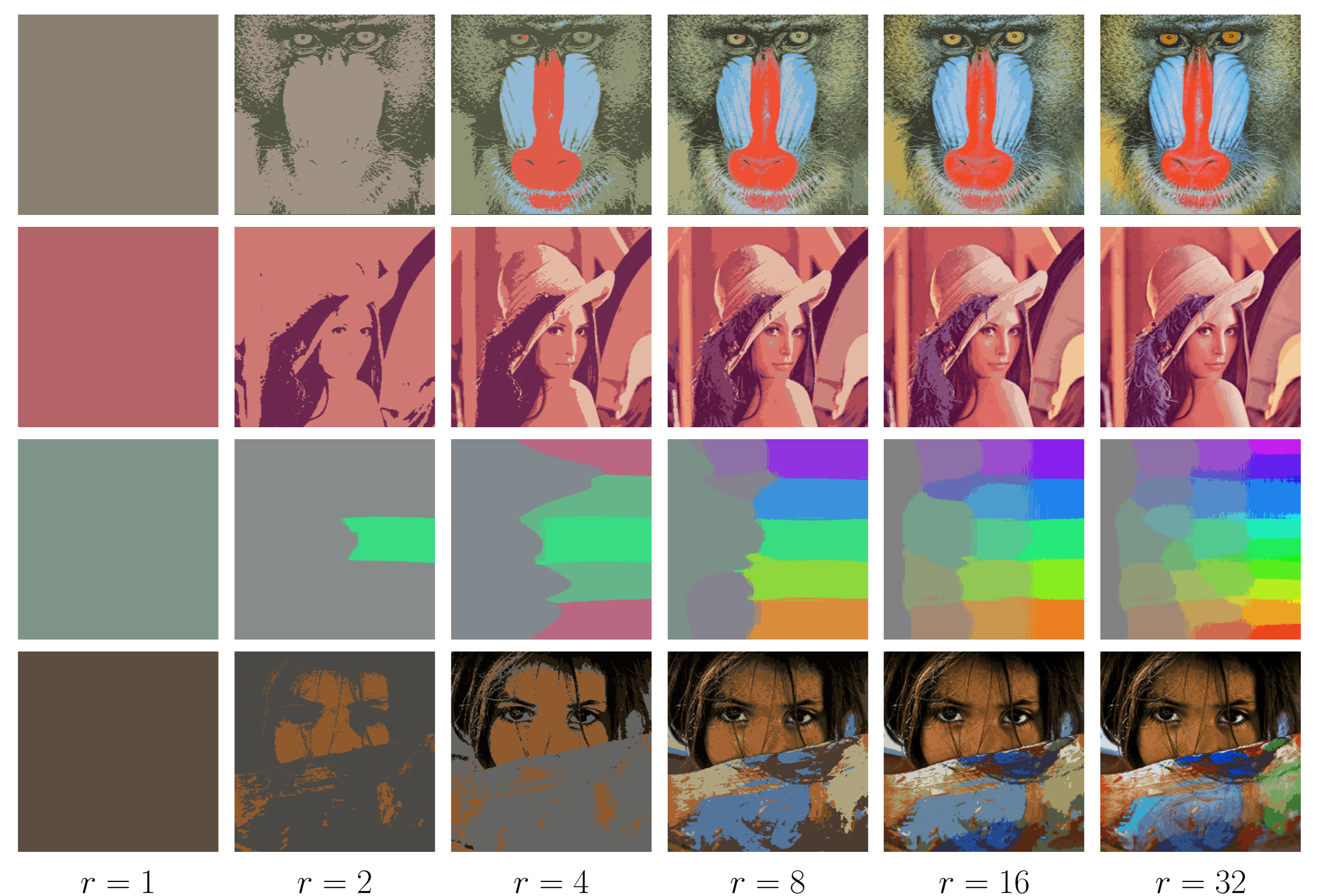
The hierarchical mixture of exponential families contains all the resolution \rightarrow progressive representation

AUTOMATIC LEARNING OF m

The user specifies a minimum mixture quality (KLD) between the source density f and the simplified mixture. Based on the hierarchical mixture of exponential families, a standard binary search on the resolution allows us to quickly find the optimal number of components in the simplified mixture.

EXPERIMENTS

- Application to clustering-based image segmentation.
- Initial mixtures contain 32 components.
- Simplification and the visual simplification qualities increase with the resolution.
- Automatic learning of *optimal* value of m (KLD min = 0.2, vary according to image semantic):
 - Baboon: $m = 10$
 - Lena: $m = 14$
 - Colormap: $m = 16$
 - Shantytown: $m = 23$



CONCLUSION

In this paper, we proposed a new algorithm which

1. creates a hierarchical representation families of a given mixture of exponential families f ,
2. provides a progressive representation of f ,
3. and learns the *optimal* number of components m in the simplified mixture.

jMEF: Open source Java library for Mixture of Exponential Families available on-line at: www.lix.polytechnique.fr/~nielsen/MEF