K-MLE: A FAST ALGORITHM FOR LEARNING STATISTICAL MIXTURE MODELS

Frank Nielsen

Sony Computer Science Laboratories, Inc. Japan

ABSTRACT

We present a fast and generic algorithm, k-MLE, for learning statistical mixture models using maximum likelihood estimators. We prove theoretically that k-MLE is dually equivalent to a Bregman k-means for the case of mixtures of exponential families (e.g., Gaussian mixture models). k-MLE is used to initialize appropriately the expectation-maximization algorithm. We also show experimentally that k-MLE outperforms the EM technique with standard initialization by considering modeling color images using high-dimensional Gaussian mixture models.

Index Terms— Gaussian mixtures, exponential families, Bregman divergences, maximum likelihood estimation

1. INTRODUCTION AND BACKGROUND

Statistical mixture models are commonly used in signal processing [1]. To sample from a finite parametric mixture model with mixture density $p(x|w_1, \theta_1, ..., w_k, \theta_k) = \sum_{i=1}^k w_i p(x|\theta_i)$ (with $\forall i, w_i > 0$ and $\sum_{i=1}^k w_i = 1$), we first draw at random the component i from which the sample emanates (using a k-nomial distribution based on the weights w_i's [1]), and then sample the observation from the chosen component distribution $p(x|\theta_i)$. For example, Gaussian mixture models (GMMs) have component parameters $\theta_i = (\mu_i, \Sigma_i)$, the means and symmetric covariance matrices. Let D denote the parameter dimension ($D = d + \frac{d(d+1)}{2} = \frac{d(d+3)}{2}$ for Gaussian distributions). The mixture model is thus defined by $c = (k-1) + kD = k(D+1) - 1$ parameters. Let Θ denote the set of mixture parameters $w_1, \theta_1, ..., w_{k-1}, \theta_{k-1}, \theta_k$ (with $w_k = 1 - \sum_{i=1}^{k-1} w_i$).

In practice, given a set X of n independently and identically distributed (iid.) observations $x_1, ..., x_n$ drawn from a statistical mixture model, one needs to *estimate* all the mixture parameters Θ. The major obstacle is that we are *missing* the component labels z_i's from which the x_i's have been sampled from. The maximum likelihood estimator (MLE) infers the parameters by maximizing the complete likelihood function (or equivalently its log-likelihood function) that measures the fitting quality of a mixture model given the prescribed observations:

$$L(x_1, ..., x_n|\Theta) = \prod_{i=1}^n p(x_i|\Theta) = \prod_{i=1}^n \prod_{j=1}^k p^{\delta_j(z_i)}(x_i|\theta_j),$$

where $\delta_j(z_i)$ is the indicator function returning 1 if and only if x_i has been sampled from the j-th component (ie., $z_i = j$), and 0 otherwise. Since there are k^n possible labels for the n observations, it is not tractable to *globally optimize* the likelihood function with the missing hidden variables z_i's. To overcome this optimization, one traditionally uses the *expectation-maximization* scheme [1] (EM). EM is a *soft clustering* technique [2] which locally converges to a *local maximum* of the log-likelihood function:

$$l(x_1, ..., x_n|\Theta) = \sum_{i=1}^n \sum_{j=1}^k \delta_j(z_i) \log p(x_i|\theta_j). \quad (1)$$

The expectation-maximization algorithm proceeds after initialization iteratively as follows:

Expectation (E). Compute the $n \times k$ weight membership matrix $W = [w_{ij}]$ using Bayes' rule [1]:
$w_{ij} = \Pr(z = j|x_i, \Theta) = \frac{w_j p(x_i|\theta_j)}{\sum_{j=1}^k w_j p(x_i|\theta_j)}$.

Maximization (M). Update the mixture parameters:
$w_j = \frac{1}{n} \sum_{i=1}^n w_{ij}$, and
$\theta_j = \arg\max \sum_{i=1}^n \log p(x_i|\theta) p(j|x_i)$.

In case of GMMs, the M-step [1] becomes $\mu_j = \frac{\sum_{i=1}^n w_{ij} x_i}{\sum_{i=1}^n w_{ij}}$ and $\Sigma_j = \frac{\sum_{i=1}^n w_{ij}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum_{i=1}^n w_{ij}}$. EM is monotonically increasing the log-likelihood function and is guaranteed to converge to a local optimum [1]. In practice, one needs to stop EM when the log-likelihood improvement becomes smaller than a prescribed threshold (otherwise EM iterates forever). Since EM is locally converging, we usually run EM with different initial configurations to track the best results. See Zhang et al. [3] for a recent efficient pruning technique to accelerate the multiple restart EM. The complexity of learning the best likelihood GMM has been recently addressed in a series of papers (see [4]). In this work, we propose a fast method to learn a mixture model, or to initialize purposely EM.

`www.informationgeometry.org/k-MLE/`

The paper is organized as follows: Section 2 describes the k-MLE algorithm. Section 3 refines the algorithm for mixtures of exponential families (including GMMs), and reinterpret it as a particular Bregman clustering [5]. Section 4 reports on the experimental performance of k-MLE and discusses on its initialization. Finally, Section 5 wraps up the contribution.

2. K-MAXIMUM LIKELIHOOD ESTIMATION

Although EM is widely used to estimate mixture models [1], it is *counter-intuitive* from the modeling point of view to deal with soft memberships for the latent variables z_i's. Soft clustering is rather a mathematical convenience for the EM optimization [1] to improve the likelihood function. We prefer to stick to the sampling *hard* membership property of observations, and propose the following iterative algorithm: k-MLE.

Initialize. Initialize the mixture components θ_i's *distinctively* (ie., $\forall i, j \in \{1, ..., k\}, j \neq i, \theta_i \neq \theta_j$).

Iterate until convergence.

Assign. Assign observations x_i to their *most likely* component (hard membership) using the current parameter estimates (and without considering the mixture weights). This yields a partition of $x_1, ..., x_n$ into k clusters: $X = \uplus_{i=1}^{k} \mathcal{C}_i$.

Re-estimate. For each cluster, re-estimate the component parameters using maximum likelihood estimators: $\forall i \in \{1, ..., k\}, \theta_i = \arg\max_\theta l(\mathcal{C}_i, \theta)$.

Finalize. Set $w_i = \frac{|\mathcal{C}_i|}{n}$ to the *proportion* of observations belonging to the i-th cluster component $\mathcal{C}_i$.

The next section considers the case of mixtures of exponential families that includes common statistical models: eg., GMMs, Rayleigh mixture models (RMMs for ultrasound imagery), Beta/Dirichlet mixtures, etc.

3. K-MLE: EXPONENTIAL FAMILY CASE

3.1. Maximum likelihood estimation

We recall the closed-form formula [6] of the MLE when dealing with exponential families, and introduce the Legendre duality. Let $x_1, ..., x_n$ be n iid. sample from an exponential family distribution [5, 6] $X \sim p_F(x; \theta) = \exp(\langle t(x), \theta \rangle - F(\theta) + k(x))$, where $\langle x, y \rangle$ denotes the inner product, $t(x)$ denotes the sufficient statistic, θ the natural parameter, $k(x)$ the auxiliary carrier measure and F the log-normalizer [6, 5]. Exponential families are log-linear models since $\log p_F(x; \theta) = \langle t(x), \theta \rangle - F(\theta) + k(x)$. For example, the multivariate Gaussians density [6] $p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\Sigma|}} \exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$ canonically decomposes as $t(x) = (x, -xx^T)$, $k(x) = 0$, $F(\theta, \Theta) = \frac{1}{4}\theta^T\Theta^{-1}\theta - \frac{1}{2}\log|\Theta| + \frac{d}{2}\log\pi$, see [6].

The log-likelihood function is $l_n(x_1, ..., x_n; \theta) = \sum_{i=1}^{n}(\langle t(x_i), \theta \rangle - F(\theta) + k(x_i))$. Removing all the $k(x_i)'s$ terms independent of θ, maximizing the log-likelihood function wrt. θ amounts to maximize $\sum_{i=1}^{n}(\langle t(x_i), \theta \rangle - F(\theta)$. It follows that

$$\nabla F(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} t(x_i). \quad (2)$$

The minimum is unique for exponential families since F is convex (the Hessian $\nabla^2 F$ is positive definite). For Gaussians, we have $\nabla F(\theta) = (\mu, -(\Sigma + \mu\mu^T))$ [6]. It follows that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T - \hat{\mu}\hat{\mu}^T$. (Note that in this case, the MLE is biased [1]. The unbiased covariance matrix is $\frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T - \hat{\mu}\hat{\mu}^T$.)

Since F is convex, we can also prove the "closed-form" of Eq. 2 using its Legendre conjugate F^*: The Legendre transformation is defined by $F^*(y) = \max_\theta \langle y, \theta \rangle - F(\theta)$. The maximum is obtained for $y - \nabla F(\theta) = 0$, that is $\theta = (\nabla F)^{-1}(y)$. Thus we have $F^*(\nabla F(\theta)) = \langle \nabla F(\theta), \theta \rangle - F(\theta)$. Let $\eta = \nabla F(\theta)$ (ie., $\theta = (\nabla F)^{-1}(\eta) = \nabla F^*(\eta)$). It follows that the log-likelihood maximization amounts to maximize $\max_\theta \sum_{i=1}^{n}(\langle t(x_i), \theta \rangle - F(\theta)$. We rewrite the equation as $\max_\theta \sum_{i=1}^{n} F^*(\eta) + \langle t(x_i) - \eta, \nabla F^*(\eta) \rangle - F^*(t(x_i)) + F^*(t(x_i))$, and introduce the dual Bregman divergence $B_{F^*}(x : y) = F^*(x) - F^*(y) - \langle x - y, (\nabla F^*)(y) \rangle$ to have equivalently $\max_\theta \sum_{i=1}^{n} -B_{F^*}(t(x_i) : \eta) + F^*(t(x_i))$. That is equivalent to minimize $\min_\theta \sum_{i=1}^{n} B_{F^*}(t(x_i) : \eta)$. For Gaussian distributions, the convex conjugate of the log-normalizer F is $F^*(\eta, H) = -\frac{1}{2}\log(1 + \eta^T H^{-1}\eta) - \frac{1}{2}\log|-H| - \frac{d}{2}\log 2\pi e$, see [6]. Thus maximizing the log-likelihood is equivalent to minimizing the average Bregman divergence [5] for the dual convex conjugate $\min_\theta \sum_{i=1}^{n} B_{F^*}(t(x_i) : \nabla F(\theta))$. This right-sided minimization is always *independent* of the generator [5, 7] and yields $\nabla F(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} t(x_i) = \hat{\eta}$ or $\hat{\theta} = \nabla F^*(\frac{1}{n}\sum_{i=1}^{n} t(x_i))$.

3.2. k-MLE as Bregman k-means

Banerjee et al. [5] showed that there exists a bijection between exponential families and Bregman divergences:

$$\log p_F(x; \theta) = -B_{F^*}(t(x) : \eta) + k(x) + F^*(t(x))$$

with $\eta = \nabla F(\theta)$ denoting the dual moment parameter. It follows that k-MLE for computing a mixture of k exponential families with log-normalizer F on $x_1, ..., x_n$ is equivalent to a Bregman k-means [5] on the sufficient statistic set $y_1 = t(x_1), ..., y_n = t(x_n)$ for the dual convex conjugate F^*. Indeed, by removing all terms $k(x_i)$'s and $F^*(t(x_i))$'s independent of Θ, we can rewrite the k-MLE optimization of the

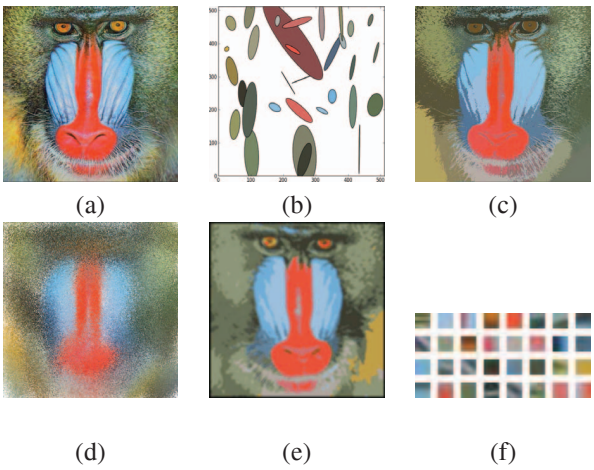

Fig. 1. Modeling a color image using a Gaussian mixture model (GMM): (a) `Baboon` source image, (b) a 5D 32-GMM modeling, (c) hard segmentation using the GMM, (d) sampling the 5D GMM, (e) Mean colors (8×8 patches) for GMM with patch size $s = 8$, (f) hard segmentation for $s = 8$ patch size.

log-likelihood function of Eq. 1 as

$$\max_{\Theta} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_j(z_i) \log p_F(x_i|\theta_j)$$

$$\min_{\eta} \sum_{i=1}^{n} \sum_{j=1}^{k} \delta_j(z_i)(B_{F^*}(t(x_i) : \eta_j) - k(x_i) - F^*(t(x_i))$$

$$\equiv \min_{\eta} \sum_{i=1}^{n} \min_{j=1}^{k} B_{F^*}(t(x_i) : \eta_j)$$

This immediately gives a proof of the monotonous convergence of k-MLE from the convergence of Bregman k-means [5]. Note that it is already known that k-means [8] can be interpreted as a hard version of the EM algorithm for a mixture of spherical Gaussians [9]. We extend this interpretation to a broader setting: k-MLE is dually equivalent to a "sufficient" Bregman k-means. In small dimensions, we can speed-up the assign step of k-MLE by using the vantage point tree proximity location data structure [10]. Note that for the case Gaussian distributions, the component distributions $p_F(x; \mu_i, \Sigma_i)$ induce a partition of the space into an *anisotropic Voronoi diagram* [11].

Although we have shown that k-MLE is theoretically equivalent to a Bregman k-means on the sufficient statistic data set $Y = \{y_i = t(x_i) \mid x_i \in X\}$ for the dual Legendre conjugate function F^*, it raises two problems in practice:

- First, the dimension D of the sufficient statistic space $\mathbb{Y}$ maybe much higher than the dimension d of the original space $\mathbb{X}$. (For example, $D = \frac{d(d+3)}{2} = \dim \mathbb{Y}$ for Gaussians instead of $d = \dim \mathbb{X}$).
- Second, the Legendre convex conjugate F^* may not be expressed in closed-form (eg., mixtures of Beta/Dirichlet distributions).

Thus it is worth working on the *primal* space $\mathbb{X}$ using k-MLE on X rather than the dual Bregman k-means on $Y \in \mathbb{Y}$.

Banerjee et al. [5] showed that EM on exponential families amount to a soft Bregman clustering problem. We fill the gap by considering the hard membership clustering k-MLE for estimating mixture models. We initialize the soft Bregman clustering with k-MLE. Since both hard clustering k-MLE and soft clustering EM monotonically converge, we compare their efficiency for a standard initialization. Let us observe that k-MLE always converges after a *finite* number of iterations but EM requires to fix a prescribed stopping criterion as it always keep improving the log-likelihood.

4. EXPERIMENTAL RESULTS

We consider modeling color images using GMMs [12]. For each pixel p_i of the image (Figure 1(a)), we associate a 5D xyRGB point $p_i = (x_i, y_i, r_i, g_i, b_i)$ by stacking the (x, y) pixel coordinates with its red green and blue color attributes (r_i, g_i, b_i). A color image of width w and height h is thus handled as a corresponding set of $n = w \times h$ 5D points $p_1, ..., p_n$. We then learn a 5D GMM with $k = 32$ components (Figure 1(b)).

We can segment the image by assigning to each pixel the mean color of the component that gives the highest probability (Figure 1(c)). We can also use the *generative statistical mixture model* to sample an "observation image" by drawing xyRGB points from the 5D GMM (Figure 1(d)). We observe that in the sample image, we loose high-frequencies (edges) (see Figure 1(d)). We build up a Laplacian image pyramid [13] and consider the smallest resolution Gaussian image that has been high-frequency band-passed filtered out. We may also consider at each pixel position of the color image, a *patch* of side length s instead of a single color pixel. In that case, we transform the source image into a point set in dimension $d = 2 + 3s^2$ by stacking the (x, y) pixel coordinates with the color information of the patch. Figure 1(f) displays the mean colors μ_i for a 32-GMM with patch size 8×8. Although k-MLE is used to better initialize EM, we compared experimentally hard k-MLE and soft EM on a benchmark set of color images using the *same* initialization parameters. We initialized the GMM by performing `kmeans++` [14] and for each cluster retrieving the centroid μ_i and covariance matrix Σ_i. Figure 2 shows that k-MLE (hard membership) outperforms EM (soft membership), and that both algorithms are monotonically converging as predicted by the theory.

Since k-MLE amounts to an equivalent Bregman k-means for the dual log-normalizer on the space of sufficient statistics, we can further use `bregkmeans++` [15] to initialize accordingly the distribution parameters. Let k-`MLE++` be that initial-

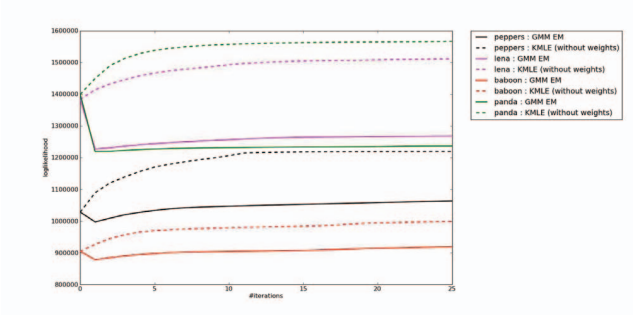

Fig. 2. Log-likelihood performance for k-MLE and expectation maximization (EM) on test images (`baboon`, `lena`, `panda`, `peppers`) using the same `kmeans++` initialization. k-MLE (dashed curves) performs better than EM (plain curves).

ization in the primal space of observations. This probabilistic initialization *guarantees* a good likelihood initialization wrt. the optimal likelihood [15]. This works well for singly-parameter exponential families (order 1), but face a degenerate situation otherwise. For example, for multivariate normals, for a chosen point y, we have $\eta = y$ yielding a degenerate MLE parameter equation: $(x, x^T x) = y = (\mu, \Sigma - \mu^T \mu)$ yielding $\Sigma = 0$, a non positive definite matrix. In general, we overcome this boundary value issue by fixing the $D - 1$ values to an arbitrary domain-valid value. (For Gaussians, say $\mu = x$ and $\Sigma = I$ the identity matrix, and k-`MLE++` amount to the regular k-`means++`.)

5. CONCLUSION

We described an efficient generic algorithm k-MLE for learning iteratively statistical mixture models. k-MLE can be used to purposely initialize the expectation-maximization technique. Although we proved its theoretical equivalence with a dual Bregman k-means [5] for the case of exponential families, it is rather practical to implement the primal k-MLE because it keeps the original low-dimensional observation space and do not require to explicitly manipulate a Legendre conjugate function that may not be in closed-form (e.g., mixture of Dirichlet distributions [1]).

Acknowledgements.
FN (5793b870) would like to thank Joris Geessels, and Dr Kitano and Dr Tokoro for their support.

6. REFERENCES

[1] G. McLachlan and D. Peel, *Finite mixture models*, Wiley-Interscience, 2000.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[3] Z.e Zhang, B. T. Dai, and A. K. H. Tung, "Estimating local optimums in EM algorithm over Gaussian mixture model," in *International Conference on Machine Learning*, 2008, pp. 1240–1247.

[4] A. Moitra and G. Valiant, "Settling the polynomial learnability of mixtures of Gaussians," in *Foundations of Computer Science*, 2010, pp. 93–102.

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

[6] F. Nielsen and V. Garcia, "Statistical exponential families: A digest with flash cards," 2009, arXiv.org:0911.4863.

[7] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2048–2059, 2009.

[8] S. P. Lloyd, "Least squares quantization in PCM," Tech. Rep., Bell Laboratories, 1957.

[9] L. Bottou and Y. Bengio, "Convergence properties of the k-means algorithms," in *Neural Information Processing Society*, 1994, pp. 585–592.

[10] F. Nielsen, P. Piro, and M. Barlaud, "Bregman vantage point trees for efficient nearest neighbor queries," in *International Conference on Multimedia & Expo*, 2009, pp. 878–881.

[11] F. Labelle and J. R. Shewchuk, "Anisotropic Voronoi diagrams and guaranteed-quality anisotropic mesh generation," in *Symposium on Computational Geometry*, 2003, pp. 191–200.

[12] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1026–1038, 2002.

[13] M. Do and M. Vetterli, "Framing pyramids," *IEEE Transactions on Signal Processing*, vol. 51, no. 9, pp. 2329–2342, 2003.

[14] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Symposium on Discrete Algorithms*, 2007, pp. 1027–1035.

[15] M. R. Ackermann and J. Blömer, "Bregman clustering for separable instances," in *Scandinavian Workshop on Algorithm Theory*, 2010, pp. 212–223.