Online k-MLE for mixture modeling with exponential families

Christophe Saint-Jean[1] and Frank Nielsen[2]

[1] Mathématiques, Image, Applications (MIA), Université de La Rochelle, France
[2] LIX, École Polytechnique, France

Abstract. This paper address the problem of online learning finite statistical mixtures of exponential families. A short review of the Expectation-Maximization (EM) algorithm and its online extensions is done. From these extensions and the description of the k-Maximum Likelihood Estimator (k-MLE), three online extensions are proposed for this latter. To illustrate them, we consider the case of mixtures of Wishart distributions by giving details and providing some experiments. **Keywords:** Mixture Modeling, Online learning, k-MLE, Wishart distribution.

1 Introduction

Mixture models are a powerful and flexible tool to model an unknown smooth probability density function as a weighted sum of parametric density functions $f_j(x;\theta_j)$:

$$f(x;\theta) = \sum_{j=1}^{K} w_j f_j(x;\theta_j), \text{ with } w_j > 0 \text{ and } \sum_{j=1}^{K} w_j = 1, \tag{1}$$

where K is the number of components of the mixture. The maximum likelihood principle is a popular approach to find the unknown parameters $\theta = \{(w_j, \theta_j)\}_j$ of f. Given $\chi = \{x_i\}_{i=1}^{N}$ a set of N independent and identically distributed observations, the maximum likelihood estimator $\hat{\theta}^{(N)}$ is defined as the maximizer of the likelihood, or equivalently of the average log-likelihood:

$$\bar{l}(\theta;\chi) = N^{-1} \sum_{i=1}^{N} \log \sum_{j=1}^{K} w_j f_j(x_i;\theta_j). \tag{2}$$

For $K > 1$, the sum of terms appearing inside a logarithm makes this optimization quite difficult.

The goal of this paper is to propose such a kind of estimator but for the online setting, that is when observations x_i are available one after another. This case appears when dealing with data streams or when data sets are large enough not to fit in memory. Ideally, online methods aim to get same convergence properties as batch ones while having a single pass over the dataset. This topic receives increasing attention due to the recent challenges associated to massive datasets.

The paper is organized as follows: Section 2 recalls the basics of Expectation-Maximization (EM) algorithm and some of its online extensions. Section 3 describes the k-MLE technique which is derived from the formalism of EM. In the same section, two online versions of k-MLE are proposed and detailed. Section 4 gives an example of the mixture of Wishart distributions and provides some experiments before concluding in Section 5.

2 A short review of online mixture learning

Before reviewing some online methods, one has to recall the basics of mixture modeling with the Expectation-Maximization (EM) algorithm [1] in the batch setting.

2.1 EM for mixture learning

Let Z_i be a categorical random variable over $1, ..., K$ whose parameters are $\{w_j\}_j$, that is, $Z_i \sim \mathrm{Cat}_K(\{w_j\}_j)$. Also, assuming that $X_i|Z_i = j \sim f_j(\cdot;\theta_j)$, the unconditional mixture distribution f in Eq. 1 is recovered by marginalizing their joint distribution over Z_i. Obviously, Z_i is a latent (unobservable) variable so that the realizations x_i of X_i (resp. (x_i, z_i) of (X_i, Z_i)) is often viewed as an incomplete (resp. complete) data observation. For convenience, we consider in the following that Z_i is a random vector $[Z_{i,1}, Z_{i,2}, \ldots, Z_{i,k}]$ where $Z_{i,j} = 1$ iff. X_i arises from the j-th component of the mixture and 0 otherwise[3]. Similarly to Eq. 2, the average complete log-likelihood function can be written as:

$$\begin{aligned}\bar{l}_c(\theta;\chi_c) &= N^{-1}\sum_{i=1}^{N}\log\prod_{j=1}^{K}\left(w_j f_j(x_i;\theta_j)\right)^{z_{i,j}}, \\ &= N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{K} z_{i,j}\log(w_j f_j(x_i;\theta_j)), \end{aligned} \tag{3}$$

where $\chi_c = \{(x_i, z_i)\}_{i=1}^N$, is the set of complete data observations. Here comes the EM algorithm which optimizes $\bar{l}(\theta;\chi)$ (proof in [1]) by repeating two steps until convergence. For iteration t:

E-Step Compute $\mathcal{Q}(\theta;\hat{\theta}^{(t)},\chi) = \mathbb{E}_{\hat{\theta}^{(t)}}[\bar{l}_c(\theta;\chi_c)|\chi]$. Since $\bar{l}_c$ is linear in $z_{i,j}$, this step amounts to compute:

$$\hat{z}_{i,j}^{(t)} = \mathbb{E}_{\hat{\theta}^{(t)}}[Z_{i,j} = 1|X_i = x_i] = \frac{\hat{w}_j^{(t)} f_j(x_i;\hat{\theta}_j^{(t)})}{\sum_{j'}\hat{w}_{j'}^{(t)} f_{j'}(x_i;\hat{\theta}_{j'}^{(t)})}. \tag{4}$$

[3] Thus, Z_i is distributed according to the multinomial law $\mathcal{M}_K(1, \{w_j\}_j)$.

M-Step Update mixture parameters by maximizing $\mathcal{Q}$ over θ (*i.e.*, Eq. 3 where hidden values $z_{i,j}$ are replaced by $\hat{z}_{i,j}^{(t)}$).

$$\hat{w}_j^{(t+1)} = \frac{\sum_{i=1}^N \hat{z}_{i,j}^{(t)}}{N}, \quad \hat{\theta}_j^{(t+1)} = \arg\max_{\theta_j \in \Theta_j} \sum_{i=1}^N \hat{z}_{i,j}^{(t)} \log\left(f_j(x_i;\theta_j)\right) \tag{5}$$

Remark that while $\hat{w}_j^{(t+1)}$ is always known in closed-form whatever f_j are, $\hat{\theta}_j^{(t+1)}$ are obtained by component-wise specific optimization involving **all** observations.

More generally, the improvement of $\bar{l}(\theta;\chi)$ is guaranteed whatever the increase of $\mathcal{Q}$ is in the M-Step. This leads to the Generalized EM algorithm (GEM) when partial maximization is performed.

2.2 Online extensions

For the online setting, it is now more appropriate to denote $\hat{\theta}^{(N)}$ the current parameter estimate instead of $\hat{\theta}^{(t)}$. In the literature, we mainly distinguish two approaches according to whether the initial structure of EM (alternate optimization) is kept or not.

The first online algorithm, due to Titterington [2], corresponds to the direct optimization of $\mathcal{Q}(\theta;\hat{\theta}^{(N)},\chi)$ using a second-order stochastic gradient ascent:

$$\hat{\theta}^{(N+1)} = \hat{\theta}^{(N)} + \gamma^{(N+1)} I_c^{-1}(\hat{\theta}^{(N)}) \nabla_\theta \log f(x_{N+1};\hat{\theta}^{(N)}), \tag{6}$$

where $\{\gamma_N\}$ is a decreasing sequence of positive step sizes ($\gamma_N = N^{-1}$ in the original paper) and the hessian $\nabla^2\mathcal{Q}$ of $\mathcal{Q}$ is approximated by the Fisher Information matrix I_c for the complete data ($I_c(\hat{\theta}^{(N)}) = -\mathbb{E}_{\hat{\theta}_j^{(N)}}[\frac{\log p(x,z;\theta)}{\partial\theta^t \partial\theta}]$). A major issue with that method is that $\hat{\theta}^{(N)}$ does not necessarily follow the parameters constraints.

This problem is coped by the approach of Cappé and Moulines [3] who proposed to replace the E-Step by a stochastic approximation step:

$$\hat{\mathcal{Q}}^{(N+1)}(\theta;\hat{\theta}^{(N)},\chi^{(N+1)}) = \hat{\mathcal{Q}}^{(N)}(\theta;\hat{\theta}^{(N)},\chi^{(N)}) + \gamma_{N+1}(\mathbb{E}_{\hat{\theta}^{(N)}}[\bar{l}_c(\theta;\{x_{N+1},z_{N+1}\})|x_{N+1}] - \hat{\mathcal{Q}}^{(N)}(\theta;\hat{\theta}^{(N)},\chi^{(N)})). \tag{7}$$

Since the M-Step remains unchanged (maximizing the function $\theta \mapsto \hat{\mathcal{Q}}^{(N+1)}(\theta)$), the constrains on parameters are automatically satisfied. This method is the starting point of our proposals. One may also mention the "Incremental EM" [4] which is not detailed here. Note that previous formalisms are not limited to mixture models.

3 Online k-Maximum Likelihood Estimator

3.1 k-MLE for mixture learning

In this section, we describe the k-MLE algorithm, a faster alternative to EM, as introduced in [5]. The goal is now to maximize directly $\bar{l}_c(\theta; \chi_c)$. In the above description of EM, value $\hat{z}_{i,j}^{(t)}$ may be interpreted as a soft membership of x_i to the j-th component of the mixture. More generally, all values $\hat{z}_{i,j}^{(t)}$ represent a soft partition of χ which may be denoted by $\hat{Z}^{(t)}$. For fixed values of θ, the partition which maximizes $\bar{l}_c$ is a strict one:

$$\max_{Z} \bar{l}_c(\theta; \chi_c) = N^{-1} \sum_{i=1}^{N} \max_{j=1}^{K} \log \left(w_j f_j(x_i; \theta_j) \right). \tag{8}$$

Doing such a maximization (also called C-Step in Classification EM algorithm [6]) after the E-Step in EM induces a split of χ into K subsets ($\chi = \bigsqcup_{j=1}^{K} \hat{\chi}_j^{(t)}$). Later on, note $\tilde{z}_{i,j}^{(t)}$ the hard membership of x_i at iteration t. Then, for a fixed optimal partition, the M-step is simpler:

$$\hat{w}_j^{(t+1)} = \frac{|\hat{\chi}_j^{(t)}|}{N}, \quad \hat{\theta}_j^{(t+1)} = \arg\max_{\theta_j \in \Theta_j} \sum_{x \in \hat{\chi}_j^{(t)}} \log f_j(x; \theta_j) \tag{9}$$

The gain in computation time is obvious since a weighted MLE involving all observations is replaced by an unweighted MLE for each subset $\hat{\chi}_j^{(t)}$. The algorithm is described in Alg. 1.

Algorithm 1: k-MLE (Lloyd's batch method)

Input: A sample $\chi = \{x_1, x_2, ..., x_N\}$
Output: Estimate $\hat{\theta}$ of mixture parameters

A good initialization for $\hat{\theta}^{(0)}$ (see [5]); t = 0;
repeat
  Partition $\chi = \bigsqcup_{j=1}^{K} \hat{\chi}_j^{(t)}$ according to $\log \hat{w}_j^{(t)} f_j(x_i; \hat{\theta}_j^{(t)})$; `// max. w.r.t. Z`
  foreach $\chi_j^{(t)}$ **do**
    $\hat{w}_j^{(t+1)} = N^{-1} |\hat{\chi}_j^{(t)}|$; `// max. w.r.t.` w_j`'s`
    $\hat{\theta}_j^{(t+1)} = \arg\max_{\theta_j \in \Theta_j} \sum_{x \in \hat{\chi}_j^{(t)}} \log f_j(x; \theta_j)$; `// max. w.r.t` θ_j`'s`
  t = t + 1;
until *Convergence of the complete likelihood*;

3.2 Proposed online extensions

In order keep ideas from online EM (stochastic E-Step) and from k-MLE (hard partition), the only possible modifications concern the assignment $z_{N+1}^{(N)}$ of the new observation x_{N+1}.

1. Online k-MLE: The most obvious heuristic is to maximize the complete log-likelihood for x_{N+1}. Indeed, unless all data is kept in memory, previous assignments for past observations are fixed. Note that these assignments are computed in order with mixture parameters $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(N-1)}$. This leads to the following rule:

$$\tilde{z}_{i,j}^{(i)} = 1 \text{ if } j = \underset{j'=1..K}{\arg\max} \log(\hat{w}_{j'}^{(i-1)} f_{j'}(x_i; \hat{\theta}_{j'}^{(i-1)})) \text{ and 0 otherwise.} \tag{10}$$

Clearly, this choice leads to a method which is similar to the Online CEM algorithm [7]. Under the assumption that components are modeled by isotropic gaussian, the MacQueens single-point iterative k-means [8] is also recovered.

2. Online Stochastic k-MLE: It is well-known that the strict partitioning can give poor results in the batch setting when mixture components are not well separated. This suggests to relax the strict maximisation and replace it by a sampling from the multinomial distribution

$$\tilde{z}_i^{(i)} \text{ sampled from } \mathcal{M}_K(1, \{p_j = \log(\hat{w}_j^{(i-1)} f_j(x_i; \hat{\theta}_j^{(i-1)}))\}_j). \tag{11}$$

Same kind of strategy was used in the Stochastic EM algorithm [6].

3. Online Hartigan k-MLE: Analogously to the Hartigan's version of k-MLE [9], one can select among all possible assignments of x_{N+1} the one which maximizes its complete likelihood **after** the M-step:

$$\tilde{z}_{i,j}^{(i)} = 1 \text{ if } j = \underset{j'=1..K}{\arg\max} \log(\hat{w}_{j'}^{(i)} f_{j'}(x_i; \theta_{j'}^{(i)})) \text{ and 0 otherwise.} \tag{12}$$

Obviously, this heuristic induces a computational overhead since K M-Steps have to be done.

To be useful, these methods require to be able to efficiently compute the MLE for components parameters. In the following, we give details for the case where these components belong to a (regular) exponential family (EF):

$$f_j(x; \theta_j) = \exp\left\{\langle t(x), \theta_j \rangle + k(x) - F(\theta_j)\right\},$$

with $t(x)$ the sufficient statistic, θ_j the natural parameter, $k(\cdot)$ the carrier measure and F the log-normalizer [10]. Under this assumption, the probability density function $p(x, z; \theta)$ is an EF[4] which can be written for the i-th observation

[4] The multinomial distribution is also an exponential family.

as:

$$\log p(x_i, z_i; \theta) = \sum_{j=1}^{K} \langle z_{i,j}, \log w_j \rangle + \sum_{j=1}^{K} \langle z_{i,j} t(x_i), \theta_j \rangle + \sum_{j=1}^{K} z_{i,j} k(x_i) - \sum_{j=1}^{K} z_{i,j} F(\theta_j) \quad (13)$$

Taking into account the summation constraint for w_j's, the M-Step reduces to simple update formulas:

$$\hat{w}_j^{(N+1)} = (N+1)^{-1} \sum_{i=1}^{N+1} \tilde{z}_{i,j}^{(i-1)}, \quad (14)$$

$$\hat{\eta}_j^{(N+1)} = (\sum_{i=1}^{N+1} \tilde{z}_{i,j}^{(i-1)})^{-1} \sum_{i=1}^{N+1} \tilde{z}_{i,j}^{(i-1)} t(x_i), \quad (15)$$

where $\eta_j = \nabla F(\theta_j)$ is the expectation parameter for the j-th component (see details in [10]). Remark that these formulas can be easily turned into recursive ones:

$$\hat{w}_j^{(N+1)} = \hat{w}_j^{(N)} + (N+1)^{-1} \left(\tilde{z}_{N+1,j}^{(N)} - \hat{w}_j^{(N)} \right), \quad (16)$$

$$\hat{\eta}_j^{(N+1)} = \hat{\eta}_j^{(N)} + (N+1)^{-1} \left(\tilde{z}_{N+1,j}^{(N)} t(x_{N+1}) - \hat{\eta}_j^{(N)} \right). \quad (17)$$

Clearly, one can recognize a step of the stochastic gradient ascent method in the expectation parameter space. Note that the functional reciprocal ∇F^{-1} must be computable to get back into natural parameter space. Algorithm 2 summarizes the Online Stochastic k-MLE.

4 Example: Mixture of Wishart distributions

4.1 Wishart distribution is a canonical (curved) exponential family

The (central) Wishart distribution [11] is the multidimensional version of the chi-square distribution and it characterizes empirical scatter matrix estimator for the multivariate gaussian distribution $\mathcal{N}_d(\mathbf{0}, S)$. Its density function can be decomposed as

$$\mathcal{W}_d(X; \theta_n, \theta_S) = \exp \left\{ \langle \theta_n, \log |X| \rangle_{\mathbb{R}} + \langle \theta_S, -\frac{1}{2} X \rangle_F + k(X) - F(\theta_n, \theta_S) \right\} \quad (18)$$

where $(\theta_n, \theta_S) = (\frac{n-d-1}{2}, S^{-1})$, $t(X) = (\log |X|, -\frac{1}{2} X)$, $k(X) = 0$ and

$$F(\theta_n, \theta_S) = \left(\theta_n + \frac{(d+1)}{2} \right) (d \log(2) - \log |\theta_S|) + \log \Gamma_d \left(\theta_n + \frac{(d+1)}{2} \right),$$

where $\Gamma_d(y) = \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma \left(y - \frac{j-1}{2} \right)$ is the multivariate gamma function defined on $\mathbb{R}_{>0}$. $\langle a, b \rangle_{\mathbb{R}} = a^\top b$ denotes the scalar product and $\langle A, B \rangle_F = \mathrm{tr}(A^\top B)$ the Fröbenius inner product (with tr the matrix trace operator). Note that this decomposition is not unique.

Algorithm 2: Online Stochastic k-MLE for (curved) exponential families

Input: A sample generator $G = x_1, x_2, ...$ yielding a data stream of observations, a batch algorithm B for the same problem, N_w a positive integer

Output: For each observation x_{N+1} with $N \geq N_w$ an estimate $\hat{\theta}^{(N+1)}$ of mixture parameters is yielded

`// Warm-Up-Step`

1 Get $\hat{\theta}^{(N)} = \{\hat{w}_j^{(N)}, \hat{\theta}_j^{(N)}\}_j$ from B with the N_w first observations of G;

2 $N = N_w$;

3 **foreach** *component j in mixture* **do** $\hat{\eta}_j^{(N)} = \nabla F(\hat{\theta}_j^{(N)})$;

4 **foreach** *new value x_{N+1} from G* **do**

5 $\quad \tilde{z}_{N+1}^{(N)}$ sampled from $\mathcal{M}_K(1, \{p_j = \log(\hat{w}_j^{(N)} f_j(x_{N+1}; \hat{\theta}_j^{(N)}))\}_j)$;

6 $\quad$ **foreach** *component j in mixture* **do**

7 $\quad\quad \hat{w}_j^{(N+1)} = \hat{w}_j^{(N)} + (N+1)^{-1}\left(\tilde{z}_{N+1,j}^{(N)} - \hat{w}_j^{(N)}\right)$;

8 $\quad\quad \hat{\eta}_j^{(N+1)} = \hat{\eta}_j^{(N)} + (N+1)^{-1}\left(\tilde{z}_{N+1,j}^{(N)} t(x_{N+1}) - \hat{\eta}_j^{(N)}\right)$;

9 $\quad$ yield mixture parameters $\hat{\theta}^{(N+1)} = \{\hat{w}_j^{(N+1)}, \hat{\theta}_j^{(N+1)} = (\nabla F)^{-1}(\hat{\eta}_j^{(N+1)})\}_j$;

10 $\quad$ N = N +1;

4.2 Details for the M-Step

Recall that find the MLE amounts to compute $(\nabla F)^{-1}$ on the average of sufficient statistics. In this specific case, the following system has to be inverted to get values of (θ_n, θ_S) given (η_n, η_S):

$$d\log(2) - \log|\theta_S| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right) = \eta_n, \tag{19a}$$

$$-\left(\theta_n + \frac{(d+1)}{2}\right)\theta_S^{-1} = \eta_S, \tag{19b}$$

where Ψ_d the derivative of the $\log \Gamma_d$ should be inverted. As far as we know, no closed-form solution exists but it can be easily solve numerically:

- Isolate θ_S in Eq. 19b: $\theta_S = \left(\theta_n + \frac{(d+1)}{2}\right)(-\eta_S)^{-1}$
- Plug it in Eq. 19a and solve numerically the following one dimensional problem:

$$d\log(2) - d\log\left(\theta_n + \frac{(d+1)}{2}\right) + \log|-\eta_S| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right) - \eta_n = 0 \tag{20}$$

  with any root-finding method on $]d-1, +\infty[$.
- Substitute the solution into Eq. 19b and solve the value for θ_S.

Whole process gives the $(\nabla F)^{-1}$ function mentioned in line 9 of Alg. 2.

Line 8 of Alg. 2 amounts to compute the following update formulas:

$$\hat{\eta_{n}}_j^{(N+1)} = \hat{\eta_{n}}_j^{(N)} + (N+1)^{-1}\left(\tilde{z}_{N+1,j}^{(N)} \log|X_{N+1}| - \hat{\eta_{n}}_j^{(N)}\right), \tag{21}$$

$$\hat{\eta_{S}}_j^{(N+1)} = \hat{\eta_{S}}_j^{(N)} - (N+1)^{-1}\left(\tilde{z}_{N+1,j}^{(N)} \tfrac{1}{2} X_{N+1} + \hat{\eta_{S}}_j^{(N)}\right). \tag{22}$$

4.3 Experiment on synthetic data-sets

In this section, we provide a preliminary empirical analysis of our proposed methods. The protocol is the following: pick a random Wishart mixture for $K = 3$ components (left) or $K = 10$ components (right), compute the Kullback-Leibler divergence between the "true" mixture and the one yielded every iteration using a Monte Carlo approximation (10^4 samples). The initialization mixture $\hat{\theta}^{(0)}$ is computed with k-MLE for the first 100 observations. The simulations are repeated 30 times for the Online Stochastic k-MLE so that it is possible to compute mean, min, max and the first and third quartiles. Also, results of online EM are reported.

From Fig. 1, one can observe a clear hierarchy between the algorithms especially when $K = 10$. One may guess that this dataset corresponds to the case when clusters components are overlapping more. Thus, the soft assignment in online EM outperforms other methods with an additional computational cost (i.e all sufficient statistics and cluster parameters have to be updated). The proof of convergence of Online Stochastic k-MLE still remain to be done while the reader may refer to the section 3.5 of the article [7] for Online k-MLE.

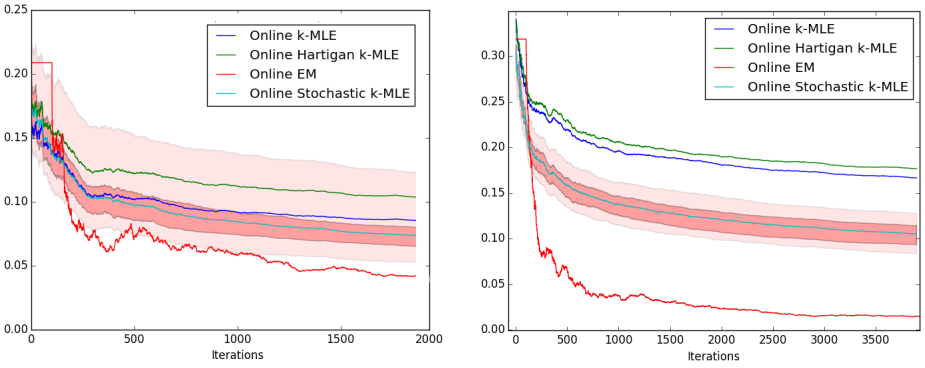

Fig. 1. $\mathrm{KL}(f(\cdot;\theta^{true})||f(\cdot;\hat{\theta}^{(N)})$ for $K = 3$ (left) and $K = 10$ (right)

5 Conclusion

This paper addresses the problem of online learning of finite statistical mixtures with a special focus on curved exponential families. The proposed methods are

fast since they require only one pass over the data stream. Further speed increase may be achieved by using distributed computing for partial sums of sufficient statistics (see [12]).

References

1. Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.
2. Titterington, D. M. : Recursive Parameter Estimation Using Incomplete Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, Volume 46, Number 2, pp. 257–267, 1984.
3. Cappé, O., Moulines, E.: On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(3):593-613, 2009.
4. Neal, R. M., Hinton, G. E.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, Learning in graphical models, pages 355-368. MIT Press, Cambridge, 1999.
5. Nielsen, F.: On learning statistical mixtures maximizing the complete likelihood Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2014), AIP Conference Proceedings Publishing, 1641, pp. 238-245, 214.
6. Celeux, G. and Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3), pp. 315-332, 1992.
7. Samé, A., Ambroise, C., Govaert, G.: An online classification EM algorithm based on the mixture model *Statistics and Computing*, 17(3), pp. 209–218, 2007.
8. MacQueen, J.: Some methods for classification and analysis of multivariate observations *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14), 1967.
9. Saint-Jean, C., Nielsen, F.: Hartigan's method for k-MLE : Mixture modeling with Wishart distributions and its application to motion retrieval. *Geometric Theory of Information*, pp. 301–330, Springer, 2014.
10. Nielsen, F., Garcia, V.: Statistical exponential families: A digest with flash cards. *http://arxiv.org/abs/0911.4863*. Nov. 2009.
11. Wishart, J.: The generalised product moment distribution in samples from a Normal multivariate population. *Biometrika*, 20(1/2), pp. 32–52, 1928.
12. Liu, Q., Ihler, A. T.: Distributed Estimation, Information Loss and Exponential Families *Advances in Neural Information Processing Systems*, pp. 1098-1106, 2014.