A new implementation of k-MLE for mixture modeling of Wishart distributions

Christophe Saint-Jean, Frank Nielsen

▶ **To cite this version:**

Christophe Saint-Jean, Frank Nielsen. A new implementation of k-MLE for mixture modeling of Wishart distributions. Geometric Science of Information, Aug 2013, Paris, France. hal-00841987

HAL Id: hal-00841987
https://hal.science/hal-00841987

Submitted on 10 Jul 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A new implementation of k-MLE for mixture modeling of Wishart distributions

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

[1] Mathématiques, Image, Applications (MIA), Université de La Rochelle, France
[2] Sony Computer Science Laboratories, Inc., Tokyo, Japan

Abstract. We describe an original implementation of k-Maximum Likelihood Estimator (k-MLE)[1], a fast algorithm for learning finite statistical mixtures of exponential families. Our version converges to a local maximum of the complete likelihood while guaranteeing not to have empty clusters. To initialize k-MLE, we propose a careful and greedy strategy inspired by k-means++ which selects automatically cluster centers and their number. The paper gives all details for using k-MLE with mixtures of Wishart (WMMs). Finally, we propose to use the Cauchy-Schwartz divergence as a comparison measure between two WMMs and give a general methodology for building a motion retrieval system.
Keywords: Mixture Modeling, Wishart, k-MLE.

1 Introduction

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

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

By far the most common case are mixtures of Gaussians for which the Expectation-Maximization (EM) method is used for decades to estimate the parameters $\{(w_j,\theta_j)\}_j$ from the maximum likelihood principle. Many extensions aimed at overcoming its slowness and lack of robustness [2]. From the seminal work of Banerjee *et al.* [3], many methods have been generalized for the exponential families in connection with the Bregman divergences. In particular, the Bregman soft clustering provides a unifying and elegant framework for the EM algorithm. In a recent work, the k-Maximum Likelihood Estimator (k-MLE) has been proposed as a fast alternative to EM. This paper proposes several variations around the initial algorithm with a specific interest for Wishart mixtures. The paper is organized as follows: Section 2 recalls some definitions, properties of Wishart distribution and gives a MLE for it; Section 3 describes the proposed algorithm and discusses how to use it for mixtures of Wishart. In Section 4, we describe an application scenario to motion retrieval before concluding in Section 5.

2 Wishart distributions

2.1 Definition

The Wishart distribution [4] is the multidimensional version of the chi-square distribution and it characterizes empirical covariance matrix estimators for the multivariate gaussian distribution. Let $\mathbb{X}$ be a n-sample consisting in independent realizations of a random gaussian vector with d dimensions, zero mean and covariance matrix S. Then $X = {}^t\mathbb{X}\mathbb{X}$ follows a central Wishart distribution with scale matrix S and degree of freedom n (DoF), denoted by $X \sim \mathcal{W}_d(n, S)$. Its density function is:

$$W_d(X; n, S) = \frac{|X|^{\frac{n-d-1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(S^{-1}X)\right\}}{2^{\frac{nd}{2}}|S|^{\frac{n}{2}}\Gamma_d\left(\frac{n}{2}\right)},$$

where for $x > 0$, $\Gamma_d(x) = \pi^{\frac{d(d-1)}{4}} \prod_{j=1}^{d} \Gamma\left(x - \frac{j-1}{2}\right)$ is the multivariate gamma function. Let us remark immediately that this definition implies that $n > d-1$.

An exponential family is a set of probability distributions admitting the following canonical decomposition:

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

with $t(x)$ the sufficient statistic, θ the natural parameter, k the carrier measure and F the log-normalizer [5]. Wishart distribution is an exponential family since

$$W_d(X;\theta_n,\theta_S) = \exp\left\{< \theta_n, \log|X| >_{\mathbb{R}} + < \theta_S, -\frac{1}{2}X >_{HS} + k(X) - F(\theta_n,\theta_S)\right\},$$

where $(\theta_n, \theta_S) = (\frac{n-d-1}{2}, S^{-1})$, $t(X) = (\log|X|, -\frac{1}{2}X)$, $\langle , \rangle_{HS}$ denotes the Hilbert-Schmidt inner product and:

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

Note that this decomposition is not unique (see another one in [6]).

2.2 Maximum Likelihood Estimator (MLE)

The framework of exponential families gives a direct solution for finding the maximum likelihood estimator from a set of i.i.d observations $X_1, \ldots, X_N$. Indeed, the MLE $\hat{\theta}$ satisfies:

$$\nabla F(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} t(X_i), \quad (2)$$

and the main difficulty is to determine the functional reciprocal $(\nabla F)^{-1}$ of ∇F (which is also ∇F^* for F^* the convex conjugate of F). In the case of the Wishart distribution, the following system must be solved:

$$\begin{cases} d\log(2) - \log|\theta_S| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right) = \eta_n, \\ -\left(\theta_n + \frac{(d+1)}{2}\right)\theta_S^{-1} = \eta_S. \end{cases} \quad (3)$$

with η_n and η_S the expectation parameters and Ψ_d the derivative of the $\log \Gamma_d$. Unfortunately, no closed-form solution is known. Instead, as pointed out in [7], it is possible to adopt an iterative scheme that alternatively yields maximum likelihood estimate when the other parameter is fixed. This is equivalent to consider two sub-families of Wishart distribution which are also exponential families. For the sake of simplicity, natural parameterizations and sufficient statistics of the decomposition in the general case are kept.

Case n fixed ($\underline{n} = 2\theta_{\underline{n}} + d + 1$):

$$F_{\underline{n}}(\theta_S) = \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|\theta_S| + \log\Gamma_d\left(\frac{\underline{n}}{2}\right), \quad k_{\underline{n}}(X) = \frac{\underline{n}-d-1}{2}\log|X|.$$

Using classical results for matrix derivatives, Eq. 2 can be easily solved :

$$-\frac{\underline{n}}{2}\hat{\theta}_S^{-1} = \frac{1}{N}\sum_{i=1}^{N} -\frac{1}{2}X_i \implies \hat{\theta}_S = N\underline{n}\left(\sum_{i=1}^{N} X_i\right)^{-1}. \tag{4}$$

Case S fixed ($\underline{S} = \theta_{\underline{S}}^{-1}$):

$$F_{\underline{S}}(\theta_n) = \left(\theta_n + \frac{d+1}{2}\right)\log|2\underline{S}| + \log\Gamma_d\left(\theta_n + \frac{d+1}{2}\right), \quad k_{\underline{S}}(X) = -\frac{1}{2}tr(\underline{S}^{-1}X).$$

Again, Eq. 2 can be numerically solved:

$$\hat{\theta}_n = \Psi_d^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\log|X_i| - \log|2\underline{S}|\right) - \frac{d+1}{2}, \qquad \hat{\theta}_n > -1. \tag{5}$$

with Ψ_d^{-1} the functional reciprocal of Ψ_d. This latter can be computed with any optimization method on bounded domain (e.g. Brent's method). Algorithm 1 summarizes the estimate for parameters of the Wishart distribution.

Algorithm 1: An estimator for the parameters of a Wishart distribution

Input: A sample $X_1, X_2, \ldots, X_N$ of $\mathcal{S}_{++}^d$
Output: Final values of $\hat{\theta}_n$ and $\hat{\theta}_S$
Initialize $\hat{\theta}_n$ with $\hat{\theta}_n = -\frac{1}{2}$;
repeat
  Update $\hat{\theta}_S$ using Eq. 4 with $\underline{n} = 2\hat{\theta}_n + d + 1$;
  Update $\hat{\theta}_n$ using Eq. 5 with $\underline{S}$ the inverse matrix of $\hat{\theta}_S$;
until *convergence of the likelihood*;

Note that even if convergence and consistency are observed in practice, the obtained final estimates remain to be proven to be the solution of the Eq. 3 or not.

3 *k*-MLE for mixtures of Wishart distributions

In [1], the k-MLE algorithm is described with the Lloyd's method : Assign all observations to their closest cluster, update clusters' parameters and so on until convergence on parameters first and then on the complete likelihood. This method can produce empty clusters (especially when the number of clusters and the data dimension are large). It is therefore preferable to adopt a different strategy because these conditions are exactly those for mixtures of Wishart.

3.1 *k*-MLE with Hartigan's method

In this part, we give an implementation of k-MLE with Hartigan's heuristic [8]: Pick an observation and optimally reassign it to another cluster (see below). This procedure is repeated until no improving transfer can be found. Interestingly, Telgarsky and Vattani [9] have noticed that this heuristic is superior to the one of Lloyd in mentioned above cases. Alg. 2 summarizes the full algorithm including the initialization which will be discussed later.

The convergence of this algorithm is fairly easy to prove since each transfer (line 13) strictly decreases the loss function:

$$\mathrm{kmeans}_{F^*,\log w}(\{t(X_i)\}_i : \{(\eta_j, w_j)\}_j) = \frac{1}{N}\sum_{i=1}^{N} \min_{j=1}^{k} \left(B_{F^*}(t(X_i) : \eta_j) - \log w_j\right),$$

with B_{F^*} the Bregman divergence for Bregman generator F^*:

$$B_{F^*}(p : q) := F^*(p) - F^*(q) - \langle p - q, \nabla F^*(q)\rangle,$$

and $\eta_1, \ldots, \eta_k$ the moment parameterization of the clusters centers. It can proved that a minimizer of $\mathrm{kmeans}_{F^*,\log w}$ is also a maximizer of the average complete log likelihood.

Surprisingly, Hartigan's original strategy doesn't prevent here from getting empty clusters. Indeed, for a singleton cluster $\mathcal{C}_i = \{X_i\}$ we have the following property :

$$\eta_i = t(X_i), \quad B_{F^*}(t(X_i) : \eta_i) - \log w_i = 0 + \log N.$$

The condition to have an improving transfer to cluster $\mathcal{C}_j$ becomes

$$\log N > B_{F^*}(t(X_i) : \eta_j) - \log |\mathcal{C}_j| + \log N \iff B_{F^*}(t(X_i) : \eta_j) < \log |\mathcal{C}_j|.$$

There is no particular reason for this condition to be always false, especially for large datasets. Thus, in order to prevent a cluster from vanishing, it is mandatory to reject every outgoing transfer for a singleton cluster (cf. line 11).

Algorithm 2: k-MLE with Hartigan's heuristic

Input: An i.i.d sample $X_1, \ldots, X_N$, an exponential family characterized by F the log-normalizer and t(X) the sufficient statistics, $\lambda > 0$

Output:

- An exponential family mixture model

$$m(x) = \sum_{j=1}^{k} w_j p_F(x; \theta_j) \text{ where } \forall j \in \{1, 2, ..., k\}, \theta_j = \nabla F^{-1}(\eta_j)$$

- A strict partition z of sample $X_1, X_2, \ldots, X_N$

$k, \{\eta_j\}_{j=1,\ldots,k}$ = DP-k-MLE++($\{t(X_i)\}_i$, F, λ); ; `// Initialization`
for $i = 1, ..., N$ **do** $z_i = \arg\min_j (B_{F^*}(t(X_i) : \eta_j) + \log k)$; `//first assignment`
;
for $j = 1, ..., k$ **do** $w_j = \frac{|\mathcal{C}_j|}{N}$ where $|\mathcal{C}_j|$ is the cardinality of cluster $\mathcal{C}_j$;
;
repeat
  need_update_w = False;
  repeat
    done_transfer = False;
    Random permute $X_1, ..., X_N$;
    foreach X_i *such that* $|\mathcal{C}_{z_i}| > 1$ **do**
      $z_i^* = \arg\min_j (B_{F^*}(t(X_i) : \eta_j) - \log w_j)$;
      if $\frac{B_{F^*}(t(X_i):\eta_{z_i^*}) - \log w_{z_i^*}}{B_{F^*}(t(X_i):\eta_{z_i}) - \log w_{z_i}} < 1$ **then**
        done_transfer = True; need_update_w = True;
        Update η_{z_i} and $\eta_{z_i^*}$:

$$\eta_{z_i} = \frac{|\mathcal{C}_{z_i}|\eta_{z_i} - t(X_i)}{|\mathcal{C}_{z_i}| - 1}, \quad \eta_{z_i^*} = \frac{|\mathcal{C}_{z_i^*}|\eta_{z_i^*} + t(X_i)}{|\mathcal{C}_{z_i^*}| + 1}$$

        $z_i = z_i^*$; Decrement $|\mathcal{C}_{z_i}|$; Increment $|\mathcal{C}_{z_i^*}|$;
  until *done_transfer is False*;
  if *need_update_w* **then for** $j = 1, ..., k$ **do** $w_j = \frac{|\mathcal{C}_j|}{N}$;
  ;
  ;
until *need_update_w is False*;

3.2 Remarks for mixtures of Wishart distributions

The k-MLE algorithm is very generic and several details must be clarified when considering mixtures of Wishart. All computations are formulated with the dual Bregman divergence $B_{F^*}(t(X_i) : \eta_j)$ which is *a priori* unknown in the general case since we can't give an expression of F^* in section 2. We only have a closed form or a numerical approximation for $F^*_{\underline{n}}$ and $F^*_{\underline{S}}$. A possible solution is to get back to the classic formulation of the complete likelihood and replace the minimization of $B_{F^*}(t(X_i) : \eta_j) - \log w_j$ by the maximization of $\log p_F(X_i; \theta_j) + \log w_j$. Unfortunately, the computational cost increases significantly because the MLE $\hat{\theta}_{j^*}$ (line 15) and $F(\hat{\theta}_{j^*})$ have to be updated after each transfer.

As remarked in [1], each component of the mixture may have its own generator F^*_j. It is of particular interest when the number of observations in $\mathbb{X}_i$ is known. In this case, several specific strategies might be explored for both the initialization and optimization (e.g. mixing B_{F^*} and $B_{F^*_{\underline{n_j}}}$).

3.3 Initialization with DP-k-MLE++

In practice, in many applications, the number of clusters is unknown and has to be estimated (e.g. penalized likelihood, cross-validation). In this paper, we adopt a greedy approach inspired by the algorithms DP-means [10] and k-MLE++ [1]. It consists in adding a cluster every time there exists an observation X_i which contributes much in proportion to the $kmeans_{F^*}$ loss function

$$kmeans_{F^*}(\{t(X_i)\}_i : \{\eta_j\}_j) = \frac{1}{N} \sum_{i=1}^{N} \min_{j=1}^{k} B_{F^*}(t(X_i) : \eta_j) \tag{6}$$

When such points exist, a new center s is chosen with a probability depending on its contribution. This procedure is summarized in Algorithm 3.

The higher the threshold λ, the lower the number of generated clusters. In particular, the value $\frac{1}{N}$ should be considered as a reasonable minimum setting for λ. For $\lambda \geq 1$, the algorithm will simply return one cluster. Since $p_i = 0$ for already selected centers, this method guarantees all centers to be *distinct*.

4 Application to motion retrieval

In a previous work [7], we described one motion captured movement $\mathbb{X}_i$ ($n_i \times d$ matrix) by the cross-product matrix $X_i = {}^t\mathbb{X}_i\mathbb{X}_i$ and derived an EM-based clustering algorithm which takes into account known values n_i[3]. To enrich the description of a movement, it is possible to define a mixture m_i per movement

[3] Matrix $\mathbb{X}_i$ are assumed to be column centered

Algorithm 3: DP-k-MLE++

Input: A sample $\{y_1 = t(X_1), \ldots, y_N = t(X_N)\}$, F a Bregman generator, $\lambda > 0$
Output: k the number of clusters, $\{\eta_1, \ldots, \eta_k\}$, a subset of $\{y_1, \ldots, y_N\}$

Choose first seed $\eta_1 = y_j$ for j uniformly random in $\{1, 2, \ldots, N\}$;
k = 1 ;
repeat
  foreach y_i **do**
    compute $p_i = \frac{\min_{j=1}^{k} B_{F^*}(y_i : \eta_j)}{\sum_{i'=1}^{N} \min_{j=1}^{k} B_{F^*}(y_{i'} : \eta_j)}$
      where F^* is the convex conjugate of F;
  if $\exists p_i > \lambda$ **then**
    Choose next seed η_{k+1} among $y_1, y_2, \ldots, y_N$ with probability p_i;
    k = k +1;
until *all* $p_i \leq \lambda$;

$\mathbb{X}_i$. For example, we can extract subsets of successive observations of different sizes and use their cross-product matrices as inputs for k-MLE. Somehow, this approach reproduces the principle of bag-of-word paradigm successfully applied many domains. Mixture m_i can be viewed as a sparse representation of local dynamics of $\mathbb{X}_i$ through their second-order moments. The problem of comparing two movements amounts to compute a dissimilarity measure between two mixtures m and m'.

When both mixtures have a single component, an immediate solution is to consider the Kullback-Liebler divergence $\mathrm{KL}(m : m')$ for two members of the same exponential family which is also the Bregman divergence on the swapped natural parameters $B_F(\theta' : \theta)$. It is important to mention that this formula does not hold when considering different generators $F_{\underline{n}}$ and $F_{\underline{n'}}$. For general mixtures of the same exponential family, KL divergence does not admit a closed form unlike the Cauchy-Schwartz divergence

$$\mathrm{CS}(m : m') = -\log \frac{\int m(x)m'(x)dx}{\int m(x)^2 dx \int m'(x)^2 dx}.$$

Skipping some details in [11], the integral of the product of mixtures can be written as

$$\int m(x)m'(x)dx = \sum_{j=1}^{k} \sum_{j'=1}^{k'} w_j w'_{j'} \exp\left\{F(\theta_j + \theta'_{j'}) - (F(\theta_j) + F(\theta'_{j'}))\right\}.$$

Note that this expression is well defined because the natural parameter space of the Wishart distribution is a *convex cone* which implies that $F(\theta_j + \theta'_{j'})$ is finite. The expression in curly brackets can be computed from Eq. 1 without much simplification to get $CS(m : m')$ and compare two movements. Details of the implementation and results for the real dataset in [7] will be in a forthcoming technical report.

5 Concluding remarks and future work

We recalled the definition and some properties of the Wishart distributions, especially its canonical decomposition as a member of an exponential family. Setting in turn, one of the two parameters, we got two other exponential (sub)families that allow to define an estimator for parameters of the Wishart distribution. Even if the experimental convergence is observed in practice, a theoretical proof and its link to the MLE remains to be done. We proposed a new implementation of the k-MLE algorithm that follows the Hartigan's method. In order to preserve the initial number of clusters, a trivial condition must be added. We also proposed an initialization method that shares the good properties of k-MLE++ and automatically sets the number of clusters. The case of Wishart mixture models were discussed. Finally, we described an application to the comparison of motion-captured movements. This is a first step towards the medium-term building of a motion retrieval system.

References

1. Nielsen, F.: k-MLE: A fast algorithm for learning statistical mixture models. In: International Conference on Acoustics, Speech and Signal Processing. (2012) pp. 869–872
2. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions. Second edition edn. Wiley Series in Probability and Statistics. Wiley-Interscience (2008)
3. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J. Clustering with bregman divergences. Journal of Machine Learning Research (6) (2005) 1705–1749
4. Wishart, J.: The generalised product moment distribution in samples from a Normal multivariate population. Biometrika **20A**(1/2) (July 1928) pp. 32–52
5. Nielsen, F., Garcia, V.: Statistical exponential families: A digest with flash cards. http://arxiv.org/abs/0911.4863 (11 2009)
6. Ji, S., Krishnapuram, B., Carin, L.: Variational Bayes for continuous hidden Markov models and its application to active learning. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(4) (2006) 522–532
7. Hidot, S., Saint Jean, C.: An Expectation-Maximization algorithm for the Wishart mixture model: Application to movement clustering. Pattern Recognition Letters **31**(14) (2010) 2318–2324
8. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1) (1979) 100–108
9. Telgarsky, M., Vattani, A.: Hartigan's method: k-means clustering without Voronoi. In: Proc. of International Conference on Artificial Intelligence and Statistics (AISTATS). (2010) pp. 820–827
10. Kulis, B., Jordan, M.I.: Revisiting k-means: New algorithms via Bayesian nonparametrics. In: International Conference on Machine Learning (ICML). (2012)
11. Nielsen, F.: Closed-form information-theoretic divergences for statistical mixtures. In: International Conference on Pattern Recognition (ICPR). (2012) pp. 1723–1726