Chapter 11
Hartigan's Method for *k*-MLE: Mixture Modeling with Wishart Distributions and Its Application to Motion Retrieval

Christophe Saint-Jean and Frank Nielsen

Abstract We describe a novel algorithm called k-Maximum Likelihood Estimator (k-MLE) for learning finite statistical mixtures of exponential families relying on Hartigan's k-means swap clustering method. To illustrate this versatile Hartigan k-MLE technique, we consider the exponential family of Wishart distributions and show how to learn their mixtures. First, given a set of symmetric positive definite observation matrices, we provide an iterative algorithm to estimate the parameters of the underlying Wishart distribution which is guaranteed to converge to the MLE. Second, two initialization methods for k-MLE are proposed and compared. Finally, we propose to use the Cauchy-Schwartz statistical divergence as a dissimilarity measure between two Wishart mixture models and sketch a general methodology for building a motion retrieval system.

Keywords Mixture modeling · Wishart · k-MLE · Bregman divergences · Motion retrieval

11.1 Introduction and Prior Work

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 $p_j(x; \theta_j)$:

C. Saint-Jean (✉)
Mathématiques, Image, Applications (MIA), Université de La Rochelle,
17000 La Rochelle, France
e-mail: christophe.saint-jean@univ-lr.fr

F. Nielsen
Sony Computer Science Laboratories, Inc., 3-14-13 Higashi Gotanda,141-0022 Shinagawa-Ku,
Tokyo, Japan

F. Nielsen
Laboratoire d'Informatique (LIX), Ecole Polytechnique, Palaiseau Cedex, France

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_11,

$$f(x) = \sum_{j=1}^{K} w_j p_j(x; \theta_j), \quad \text{with } w_j > 0 \text{ and } \sum_{j=1}^{K} w_j = 1. \tag{11.1}$$

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 [1]. From the seminal work of Banerjee et al. [2], several 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 with mixtures of exponential families. In a recent work [3], the k-Maximum Likelihood Estimator (k-MLE) has been proposed as a fast alternative to EM for learning any exponential family mixtures: k-MLE relies on the bijection of exponential families with Bregman divergences to transform the mixture learning problem into a geometric clustering problem. Thus we refer the reader to the review paper [4] for an introduction to clustering.

This paper proposes several variations around the initial k-MLE algorithm with a specific focus on mixtures of Wishart [5]. Such a mixture can model complex distributions over the set $\mathcal{S}_{++}^d$ of $d \times d$ symmetric positive definite matrices. Data of this kind comes naturally in some applications like diffusion tensor imaging, radar imaging but also artificially as signature for a multivariate dataset (region of interest in an multispectral image or a temporal sequence of measures for several sensors).

In the literature, the Wishart distribution is rarely used for modeling data but more often in bayesian approaches as a (conjugate) prior for the inverse covariance-matrix of a gaussian vector. This justifies that few works concern the estimation of the parameters of Wishart from a set of matrices. To the best of our knowledge, the only and most related work is the one of Tsai [6] concerning MLE and Restricted-MLE with ordering constraints. From the application viewpoint, one may cite polarimetric SAR imaging [7], bio-medical imaging [8]. Another example is a recent paper on people tracking [9] which applies Dirichlet process mixture model (infinite mixture model) to the clustering of covariance matrices.

The paper is organized as follows: Sect. 11.2 recalls the definition of an exponential family (EF), the principle of maximum likelihood estimation in EFs and how it is connected with Bregman divergences. From these definitions, the complete description of k-MLE technique is derived by following the formalism of the Expectation-Maximization algorithm in Sect. 11.3. In the same section, the Hartigan approach for k-MLE is proposed and discussed as well as how to initialize it properly. Section 11.4 concerns the learning of a mixture of Wishart with k-MLE. For this purpose, a iterative procedure that converges to the MLE when it exists. In Sect. 11.5, we describe an application scenario to motion retrieval before concluding in Sect. 11.6.

11.2 Preliminary Definitions and Notations

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

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

with $t(x)$ the sufficient statistic, θ the natural parameter, k the carrier measure and F the log-normalizer [10]. Most of commonly used distributions such Bernoulli, Gaussian, Multinomial, Dirichlet, Poisson, Beta, Gamma, von Mises are indeed exponential families (see above reference for a complete list). Later on in the chapter, a canonical decomposition of the Wishart distribution as an exponential family will be detailed.

11.2.1 Maximum Likelihood Estimator

The framework of exponential families gives a direct solution for finding the maximum likelihood estimator $\hat{\theta}$ from a set of i.i.d observations $\chi = \{x_1, \ldots, x_N\}$. Denoting $\mathcal{L}$ the likelihood function

$$\mathcal{L}(\theta;\chi) = \prod_{i=1}^{N} p_F(x_i;\theta) = \prod_{i=1}^{N} \exp\{\langle t(x_i),\theta\rangle + k(x_i) - F(\theta)\} \tag{11.2}$$

and $\bar{l}$ the average log-likelihood function

$$\bar{l}(\theta;\chi) = \frac{1}{N}\sum_{i=1}^{N} \left(\langle t(x_i),\theta\rangle + k(x_i) - F(\theta)\right). \tag{11.3}$$

It follows that the MLE $\hat{\theta} = \arg\max_{\Theta} \bar{l}(\theta;\chi)$ for θ satisfies

$$\nabla F(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N} t(x_i). \tag{11.4}$$

Recall that the functional reciprocal $(\nabla F)^{-1}$ of ∇F is also ∇F^* for F^* the convex conjugate of F [11]. It is a mapping from the expectation parameter space $\mathbf{H}$ to the natural parameter space Θ. Thus, the MLE is obtained by mapping $(\nabla F)^{-1}$ on the average of sufficient statistics:

$$\hat{\theta} = (\nabla F)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} t(x_i)\right). \tag{11.5}$$

Whereas determining $(\nabla F)^{-1}$ may be trivial for some univariate distributions like Bernoulli, Poisson, Gaussian, multivariate case is much challenging and lead to consider approximate methods to solve this variational problem [12].

11.2.2 MLE and Bregman Divergence

In this part, the link between MLE and Kullback-Leibler (KL) divergence is recalled. Banerjee et. al. [2] interpret the log-density of a (regular) exponential family as a (regular) Bregman divergence:

$$\log p_F(x;\theta) = -B_{F^*}(t(x) : \eta) + F^*(t(x)) + k(x), \tag{11.6}$$

where F^* is the convex conjugate (Legendre transform) of F. Skipping a formal definition, a Bregman divergence for a strictly convex and differentiable function $\varphi : \Omega \mapsto \mathbb{R}$ is

$$B_\varphi(\omega_1 : \omega_2) = \varphi(\omega_1) - \varphi(\omega_2) - \langle \omega_1 - \omega_2, \nabla\varphi(\omega_2)\rangle. \tag{11.7}$$

From a geometric viewpoint, $B_\varphi(\omega_1 : \omega_2)$ is the difference between the value of φ at ω_1 and its first-order Taylor expansion around ω_2 evaluated at ω_1. Since φ is convex, B_φ is positive, zero if and only if (iff) $\omega_1 = \omega_2$ but not symmetric in general. The expression of F^* (and thus of B_{F^*}) follows from $(\nabla F)^{-1}$

$$F^*(\eta) = \langle (\nabla F)^{-1}(\eta), \eta\rangle - F((\nabla F)^{-1}(\eta)). \tag{11.8}$$

In Eq. 11.6, term $B_{F^*}(t(x) : \eta)$ says how much sufficient statistic $t(x)$ on observation x is dissimilar to $\eta \in \mathbf{H}$.

The Kullback-Leibler divergence on two members of the same exponential family is equivalent to the Bregman divergence of the associated log-normalizer on swapped natural parameters [10]:

$$\mathrm{KL}(p_F(.;\theta_1)||p_F(.;\theta_2)) = B_F(\theta_2 : \theta_1) = B_{F^*}(\eta_1 : \eta_2). \tag{11.9}$$

Let us remark that B_F is always known in a closed-form using the canonical decomposition of p_F whereas B_{F^*} requires the knowledge of F^*. Finding the maximizer of the log likelihood on Θ amounts to find the minimizer $\hat{\eta}$ of

$$\sum_{i=1}^{N} B_{F^*}(t(x_i) : \eta) = \sum_{i=1}^{N} \mathrm{KL}\left(p_{F^*}(.; t(x_i))||p_{F^*}(.;\eta)\right)$$

on $\mathbf{H}$ since the two last terms in Eq. (11.6) are constant with respect to η.

11.3 Learning Mixtures with *k*-MLE

This section presents how to fit a mixture of exponential families with k-MLE. This algorithm requires to have a MLE (see previous section) for each component distribution p_{F_j} of the considered mixture. As it shares many properties with the EM algorithm for mixtures, this latter is recalled first. The heuristics Lloyd and Hartigan for k-MLE are completely described. Also, two methods for the initialization of k-MLE are proposed depending whether or not component distributions are known.

11.3.1 EM Algorithm

Mixture modeling is a convenient framework to address the problem of clustering defined as the partitioning of a set of i.i.d observations $\chi = \{x_i\}_{i=1,..,N}$ into "meaningful" groups regarding to some similarity. Consider a finite mixture model of exponential families (see Eq. 11.1)

$$f(x) = \sum_{j=1}^{K} w_j p_{F_j}(x; \theta_j), \tag{11.10}$$

where K is the number of components and w_j are the mixture weights which sum up to unity. Finding mixture parameters $\{(w_j, \theta_j)\}_j$ can be again addressed by maximizing the log likelihood of the mixture distribution

$$\mathcal{L}(\{(w_j, \theta_j)\}_j; \chi) = \sum_{i=1}^{N} \log \sum_{j=1}^{K} w_j p_{F_j}(x_i; \theta_j). \tag{11.11}$$

For $K > 1$, a sum of terms appearing inside a logarithm makes optimization much more difficult than the one of Sect. 11.2.1 ($K = 1$). A classical solution, also well suitable for clustering purpose, is to augment model with indicatory hidden vector variables z_i where $z_{ij} = 1$ iff observation x_i is generated for jth component and 0 otherwise. Previous equation is now replaced by the complete log likelihood of the mixture distribution

$$\mathcal{L}_{\mathbf{c}}(\{(w_j, \theta_j)\}_j; \{(x_i, z_i)\}_i) = \sum_{i=1}^{N} \sum_{j=1}^{K} z_{ij} \log \left(w_j p_{F_j}(x_i; \theta_j) \right). \tag{11.12}$$

This is typically the framework of the Expectation-Maximization (EM) algorithm [13] which optimizes this function by repeating two steps:

1. Compute $\mathcal{Q}(\{(w_j, \theta_j)\}_j, \{(w_j^{(t)}, \theta_j^{(t)})\}_j)$ the conditional expectation of $\mathcal{L}_{\mathbf{c}}$ w.r.t. the observed data χ given an estimate $\{(w_j^{(t)}, \theta_j^{(t)})\}_j$ for mixture parameters. This step amounts to compute $\hat{z}_i^{(t)} = \mathbb{E}_{\{(w_j^{(t)}, \theta_j^{(t)})\}_j}[z_i | x_i]$, the vector of responsibilities for each component to have generated x_i.

$$\hat{z}_{ij}^{(t)} = \frac{w_j^{(t)} p_{F_j}(x_i; \theta_j^{(t)})}{\sum_{j'} w_{j'}^{(t)} p_{F_{j'}}(x_i; \theta_{j'}^{(t)})}. \tag{11.13}$$

2. Update mixture parameters by maximizing $\mathcal{Q}$ (i.e. Eq. (11.12) where hidden values z_{ij} are replaced by $\hat{z}_{ij}^{(t)}$).

$$\hat{w}_j^{(t+1)} = \frac{\sum_{i=1}^N \hat{z}_{ij}^{(t)}}{N}, \quad \hat{\theta}_j^{(t+1)} = \underset{\theta_j \in \Theta_j}{\arg\max} \sum_{i=1}^N \hat{z}_{ij}^{(t)} \log\left(p_{F_j}(x_i; \theta_j)\right). \tag{11.14}$$

   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.

Many properties of this algorithm are known (e.g. maximization of $\mathcal{Q}$ implies maximization of $\mathcal{L}$, slow convergence to local maximum, etc...). In a clustering perspective, components are identified to clusters and values $\hat{z}_{ij}$ are interpreted as soft membership of x_i to cluster $\mathcal{C}_j$. In order to get a strict partition after the convergence, each x_i is assigned to the cluster $\mathcal{C}_j$ iff $\hat{z}_{ij}$ is maximum over $\hat{z}_{i1}, \hat{z}_{i2}, \ldots, \hat{z}_{iK}$.

11.3.2 k-MLE with Lloyd Method

A main reason for the slowness of EM is that all observations are taken into account for the update of parameters for each component since $\hat{z}_{ij}^{(t)} \in [0, 1]$. A natural idea is then to generate smaller sub-samples of χ from $\hat{z}_{ij}^{(t)}$ in a deterministic manner.[1] The simplest way to do this is to get a strict partition of χ with MAP assignment:

$$\tilde{z}_{ij}^{(t)} = \begin{cases} 1 & \text{if } \hat{z}_{ij}^{(t)} = \max_k \hat{z}_{ik}^{(t)} \\ 0 & \text{otherwise} \end{cases}.$$

When multiple maxima exist, the component with the smallest index is chosen. If this classification step is inserted between E-step and M-step, Classification EM (CEM) algorithm [14] is retrieved. Moreover, for isotropic gaussian components with fixed unit variance, CEM is shown to be equivalent to the Lloyd K-means algorithm [4]. More recently, CEM was reformulated in a close way under the name k-MLE [3] for the context of exponential families and Bregman divergences. In the following of the paper, we will refer only to this latter. Replacing $z_{ij}^{(t)}$ by $\tilde{z}_{ij}^{(t)}$ in Eq. (11.12), the criterion to be maximized in the M-step can be reformulated as

[1] Otherwise, convergence to a pointwise estimate of the parameters would be replaced by convergence in distribution of a Markov chain.

$$\tilde{\mathcal{L}}_{\mathbf{c}}(\{(w_j,\theta_j)\}_j;\{(x_i,\tilde{z}_i^{(t)})\}_i)=\sum_{j=1}^{K}\sum_{i=1}^{N}\tilde{z}_{ij}^{(t)}\log\left(w_j p_{F_j}(x_i;\theta_j)\right). \tag{11.15}$$

Following CEM terminology, this quantity is called the "classification maximum likelihood". Letting $C_j^{(t)}=\left\{x_i\in\chi|\tilde{z}_{ij}^{(t)}=1\right\}$, this equation can be conveniently rewritten as

$$\begin{aligned}\tilde{\mathcal{L}}_{\mathbf{c}}(\{(w_j,\theta_j)\}_j;\{\mathcal{C}_j^{(t)}\}_j)=&\sum_{x\in\mathcal{C}_1^{(t)}}\log\left(w_1 p_{F_1}(x;\theta_1)\right)\\&+\cdots+\sum_{x\in\mathcal{C}_K^{(t)}}\log\left(w_K p_{F_K}(x;\theta_K)\right).\end{aligned} \tag{11.16}$$

Each term leads to a separate optimization to get the parameters of the corresponding component:

$$\hat{w}_j^{(t+1)}=\frac{|\mathcal{C}_j^{(t)}|}{N},\quad \hat{\theta}_j^{(t+1)}=\underset{\theta_j\in\Theta_j}{\arg\max}\sum_{x\in\mathcal{C}_j^{(t)}}\log p_{F_j}(x;\theta_j). \tag{11.17}$$

Last equation is nothing but the equation of the MLE for the j-th component with a subset of χ. Algorithm 1 summarizes k-MLE with Lloyd method given an initial description of the mixture.

Algorithm 1: k-MLE (Lloyd method)

```
Input: A sample $\chi=\{x_1,x_2,...,x_N\}$, initial mixture parameters $\{\hat{w}_j^{(0)},\hat{\theta}_j^{(0)}\}_j$, $\{F_j\}_j$
       log-normalizers of exponential families
Output: Ending values for $\{\hat{w}_j^{(t)},\hat{\theta}_j^{(t)}\}_j$ are estimates of mixture parameters,
$\mathcal{C}_j^{(t)}$ a partition of $\chi$
t = 0;
repeat
    repeat
        // Partition $\chi$ in $K$ disjoint subsets with MAP assignment
        foreach $x_i\in\chi$ do $\bar{z}_i^{(t)}=\arg\max_j\log\hat{w}_j^{(t)}p_{F_j}(x_i;\hat{\theta}_j^{(t)})$;
        $\mathcal{C}_j^{(t)}=\{x_i\in\chi|\bar{z}_i^{(t)}=j\}$;
        // Update parameters $\{\theta_j\}_j$ with MLE ($\{w_j\}_j$ unchanged)
        foreach $j\in 1,...,K$ do $\hat{\theta}_j^{(t+1)}=\arg\max_{\theta_j\in\Theta_j}\sum_{x\in\mathcal{C}_j^{(t)}}\log p_{F_j}(x;\theta_j)$;
        t = t + 1;
    until Convergence of the classification maximum likelihood (Eq. 11.16);
    // Update mixture weights $\{w_j\}_j$
    foreach $j\in 1,...,K$ do $\hat{w}_j^{(t+1)}=|\mathcal{C}_j^{(t)}|/N$;
until Further convergence of the classification maximum likelihood (Eq. 11.16);
```

Contrary to CEM algorithm, k-MLE algorithm updates mixture weights after the convergence of the $\tilde{\mathcal{L}}_{\mathbf{c}}$ (line 8) and not simultaneously with component parameters. Despite this difference, both algorithms can be proved to converge to a local maximum of $\tilde{\mathcal{L}}_{\mathbf{c}}$ with same kind of arguments (see [3, 14]). In practice, the local maxima (and also the mixture parameters) are not necessary equal for the two algorithms.

11.3.3 k-MLE with Hartigan Method

In this section, a different optimization of the classification maximum likelihood is presented. A drawback of previous methods is that they can produce empty clusters without any mean of control. It occurs especially when observations are in a high dimensional space. A mild solution is to discard empty clusters by setting their weights to zero and their parameters to $\emptyset$. A better approach, detailed in the following, is to rewrite the k-MLE following the same principle as Hartigan method for k-means [15]. Moreover, this heuristic is preferred to Lloyd's one since it generally provides better local maxima [16].

Hartigan method is generally summarized by the sentence "Pick an observation, say x_c in cluster $\mathcal{C}_c$, and optimally reassign it to another cluster." Let us first consider as "optimal" the assignment x_c to its most probable cluster, say $\mathcal{C}_{j^*}$:

$$j^* = \arg\max_j \log \hat{w}_j^{(t)} p_{F_j}(x_c; \hat{\theta}_j^{(t)}),$$

where $\hat{w}_j^{(t)}$, and $\hat{\theta}_j^{(t)}$ denote the weight and the parameters of the j-th component at some iteration. Then, parameters of the two components are updated with MLE:

$$\hat{\theta}_c^{(t+1)} = \arg\max_{\theta_c \in \Theta_c} \sum_{x \in \mathcal{C}_c^{(t)} \setminus \{x_c\}} \log p_{F_c}(x; \theta_c) \tag{11.18}$$

$$\hat{\theta}_{j^*}^{(t+1)} = \arg\max_{\theta_{j^*} \in \Theta_{j^*}} \sum_{x \in \mathcal{C}_{j^*}^{(t)} \cup \{x_c\}} \log p_{F_{j^*}}(x; \theta_{j^*}). \tag{11.19}$$

The mixture weights $\hat{w}_c$ and $\hat{w}_{j^*}$ remain unchanged in this step (see line 9 of Algorithm 1). Consequently, $\tilde{\mathcal{L}}_{\mathbf{c}}$ increases by $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_{j^*})$ where $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j)$ is more generally defined as

$$\begin{aligned}\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j) = &\sum_{x \in \mathcal{C}_c^{(t)} \setminus \{x_c\}} \log p_{F_c}(x; \hat{\theta}_c^{(t+1)}) - \sum_{x \in \mathcal{C}_c^{(t)} \cup \{x_c\}} \log p_{F_c}(x; \hat{\theta}_c^{(t)}) - \log \frac{\hat{w}_c}{\hat{w}_j} \\ &+ \sum_{x \in \mathcal{C}_j^{(t)} \cup \{x_c\}} \log p_{F_j}(x; \hat{\theta}_j^{(t+1)}) - \sum_{x \in \mathcal{C}_j^{(t)} \setminus \{x_c\}} \log p_{F_j}(x; \hat{\theta}_j^{(t)}).\end{aligned} \tag{11.20}$$

This procedure is nothing more than a partial assignment (C-step) in the Lloyd method for k-MLE. This is an indirect way to reach our initial goal which is the maximization of $\tilde{\mathcal{L}}_{\mathbf{c}}$.

Following Telgarsky and Vattani [16], a better approach is to consider as "optimal" the assignment to cluster $\mathcal{C}_j$ which maximizes $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j)$

$$j^* = \arg\max_j \Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j). \tag{11.21}$$

Since $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_c) = 0$, such assignment satisfies $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_{j^*}) \geq 0$ and therefore the increase of $\tilde{\mathcal{L}}_{\mathbf{c}}$. As the optimization space is finite (partitions of $\{x_1, ..., x_N\}$), this procedure converges to a local maximum of $\tilde{\mathcal{L}}_{\mathbf{c}}$. There is no guarantee that $\mathcal{C}_{j^*}$ coincides with the MAP assignment for x_c.

For the k-means loss function, Hartigan method avoids empty clusters since any assignment to one of those empty clusters decreases it necessarily [16]. Analogous property will be now studied for k-MLE through the formulation of $\tilde{\mathcal{L}}_{\mathbf{c}}$ with η-coordinates:

$$\begin{aligned}&\tilde{\mathcal{L}}_{\mathbf{c}}(\{(w_j, \eta_j)\}_j; \{\mathcal{C}_j^{(t)}\}_j) \\ &= \sum_{j=1}^{K} \sum_{x \in \mathcal{C}_j^{(t)}} \left[F_j^*(\eta_j) + k_j(x) + \langle t_j(x) - \eta_j, \nabla F_j^*(\eta_j) \rangle + \log w_j \right].\end{aligned} \tag{11.22}$$

Recalling that the MLE satisfies $\hat{\eta}_j^{(t)} = |\mathcal{C}_j^{(t)}|^{-1} \sum_{x \in \mathcal{C}_j^{(t)}} t_j(x)$, dot product vanishes when $\eta_j = \hat{\eta}_j^{(t)}$ and it follows after small calculations

$$\begin{aligned}\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j) = &(|\mathcal{C}_c^{(t)}| - 1) F_c^*(\hat{\eta}_c^{(t+1)}) - |\mathcal{C}_c^{(t)}| F_c^*(\hat{\eta}_c^{(t)}) \\ &+ \ (|\mathcal{C}_j^{(t)}| + 1) F_j^*(\hat{\eta}_j^{(t+1)}) - |\mathcal{C}_j^{(t)}| F_j^*(\hat{\eta}_j^{(t)}) \\ &+ \ k_j(x_c) - k_c(x_c) - \log \frac{\hat{w}_c}{\hat{w}_j}.\end{aligned} \tag{11.23}$$

As far as F_j^* is known in closed-form, this criterion is faster to compute than Eq. (11.20) since updates of component parameters are immediate

$$\hat{\eta}_c^{(t+1)} = \frac{|\mathcal{C}_c^{(t)}| \hat{\eta}_c^{(t)} - t_c(x_c)}{|\mathcal{C}_c^{(t)}| - 1}, \ \hat{\eta}_j^{(t+1)} = \frac{|\mathcal{C}_j^{(t)}| \hat{\eta}_j^{(t)} + t_j(x_c)}{|\mathcal{C}_j^{(t)}| + 1}. \tag{11.24}$$

When $\mathcal{C}_c^{(t)} = \{x_c\}$, there is no particular reason for $\Phi^{(t)}(x_c, \{x_c\}, \mathcal{C}_j)$ to be always negative. Simplications occurring for the k-means in euclidean case (e.g. $k_j(x_c) = 0$, clusters have equal weight $w_j = K^{-1}$, etc...) do not exist in this more general case.

Thus, in order to avoid to empty a cluster, it is mandatory to reject every outgoing transfer for a singleton cluster (cf. line 8).

Algorithm 2 details k-MLE algorithm with Hartigan method when F_j^* are available. When only F_j are known, $\Phi^{(t)}(x_c, \mathcal{C}_c, \mathcal{C}_j)$ can be computed with Eq. (11.20). In this case, the computation of MLE $\hat{\theta}_j$ is much slower and is an issue for a singleton cluster. Its existence and possible solutions will be discussed later for the Wishart distribution. Further remarks on Algorithm 2:

(line 1) When all $F_j^* = F^*$ are identical, this partitioning can be understood as geometric split in the expectation parameter space induced by divergence B_{F^*} and additive weight $-\log w_j$ (weighted Bregman Voronoi diagram [17]).

(line 4) This permutation avoids same ordering for each loop.

(line 6) A weaker choice may be done here: any cluster $\mathcal{C}_j$ (for instance the first) which satisfies $\Phi^{(t)}(x_i, \mathcal{C}_{\bar{z}_i}, \mathcal{C}_j) > 0$ is a possible candidate still guaranteeing convergence of the algorithm. For such clusters, it may be also advisable to select $\mathcal{C}_j$ with maximum $\hat{z}_{ij}^{(t)}$.

(line 12) Obviously, this criterion is equivalent to local convergence of $\tilde{\mathcal{L}}_{\mathbf{c}}$.

As said before, this algorithm is faster when components parameters η_j can be updated in the expectation parameter space $\mathbf{H}_j$. But the price to pay is the memory needed to keep all sufficient statistics $t_j(x_i)$ for each observation x_i.

11.3.4 Initialization with DP-k-MLE++

To complete the description of k-MLE, it remains the problem of the initialization of the algorithm: choice of the exponential family for each component, initial values of $\{(\hat{w}_j^{(0)}, \eta_j^{(0)})\}_j$, number of components K. Ideally, a good initialization would be fast, select automatically the number of components (unknown for most applications) and provide initial mixture parameters not too far from a good local minimum of the clustering criterion. The choice of model complexity (i.e. the number K of groups) is a recurrent problem in clustering since a compromise has to be done between genericity and goodness of fit. Since the likelihood increases with K, many criteria such as BIC, NEC are based on the penalization of likelihood by a function of the degree of freedom of the model. Other approaches include MDL principle, Bayes factor or simply a visual inspection of some plottings (e.g. silhouette graph, dendrogram for hierarchical clustering, Gram matrix, etc...). The reader interested by this topic may refer to section M in the survey of Xu and Wunsch [18]. Proposed method, inspired by the algorithms k-MLE++ [3] and DP-means [19], will be described in the following.

At the beginning of a clustering, there is no particular reason for favoring one particular cluster among all others. Assuming uniform weighting for components, $\tilde{\mathcal{L}}_{\mathbf{c}}$ simplifies to

Algorithm 2: k-MLE (Hartigan method)

```
Input: Sample $\chi = \{x_1, .., x_N\}$, initial mixture parameters $\{(\hat{w}_j^{(0)}, \hat{\eta}_j^{(0)})\}_{j=1,..,K}$, $\{(t_j, F_j^*)\}_j$
       sufficient statistics and dual log-normalizers of exponential families
Output: Ending values for $\{(\hat{w}_j^{(t)}, \hat{\eta}_j^{(t)})\}_j$ are estimates of mixture parameters, $\mathcal{C}_j^{(t)}$ a
        partition of $\chi$
// Partition $\chi$ in $K$ disjoint subsets with MAP assignment
foreach $x_i \in \chi$ do $\bar{z}_i^{(0)} = \arg\min_j (B_{F_j^*}(t_j(x_i) : \hat{\eta}_j^{(0)}) - \log \hat{w}_j^{(0)})$;
foreach $j \in 1, ..., K$ do $\mathcal{C}_j^{(0)} = \{x_i \in \chi | \bar{z}_i^{(0)} = j\}$;
repeat
    done_transfer = False;
    Random permute $(x_1, \bar{z}_1^{(t)}), ..., (x_N, \bar{z}_N^{(t)})$;
    foreach $x_i \in \chi$ such that $|\mathcal{C}_{\bar{z}_i^{(t)}}^{(t)}| > 1$ do
        // Test optimal transfer for $x_i$ (see Eqs. 11.23 or 11.20)
        $j^* = \arg\min_j \Phi^{(t)}(x_i, \mathcal{C}_{\bar{z}_i^{(t)}}, \mathcal{C}_j)$;
        if $\Phi^{(t)}(x_i, \mathcal{C}_{\bar{z}_i^{(t)}}, \mathcal{C}_{j^*}) > 0$ then
            // Update clusters and membership of $x_i$
            $\mathcal{C}_{\bar{z}_i^{(t)}}^{(t+1)} = \mathcal{C}_{\bar{z}_i^{(t)}}^{(t)} \backslash \{x_i\}, \quad \mathcal{C}_{j^*}^{(t+1)} = \mathcal{C}_{j^*}^{(t)} \cup \{x_i\}, \quad \bar{z}_i^{(t+1)} = j^*$
            // Update only $\eta_{\bar{z}_i}, \eta_{j^*}$ with MLE ($\{w_j\}_j$ unchanged)
            $\hat{\eta}_{\bar{z}_i^{(t)}}^{(t+1)} = \frac{|\mathcal{C}_{\bar{z}_i^{(t)}}^{(t)}| \hat{\eta}_{\bar{z}_i^{(t)}}^{(t)} - t_{\bar{z}_i^{(t)}}(x_i)}{|\mathcal{C}_{\bar{z}_i^{(t)}}^{(t+1)}|}, \quad \hat{\eta}_{\bar{z}_i^{(t+1)}}^{(t+1)} = \frac{|\mathcal{C}_{\bar{z}_i^{(t+1)}}^{(t)}| \hat{\eta}_{\bar{z}_i^{(t+1)}}^{(t)} + t_{\bar{z}_i^{(t+1)}}(x_i)}{|\mathcal{C}_{\bar{z}_i^{(t+1)}}^{(t+1)}|}$
            done_transfer = True; t = t + 1;
    if done_transfer is True then
        // Update mixture weights $\{w_j\}_j$
        foreach $j \in 1, ..., K$ do $\hat{w}_j^{(t)} = N^{-1} |\mathcal{C}_j^{(t)}|$;
until done_transfer is False;
```

$$\mathring{\mathcal{L}}_{\mathbf{c}}(\{\theta_j\}_j; \{\mathcal{C}_j^{(t)}\}_j) = \sum_{j=1}^{K} \sum_{x \in \mathcal{C}_j^{(t)}} \log p_{F_j}(x; \theta_j) \qquad \textit{or equivalently to} \tag{11.25}$$

$$\mathring{\mathcal{L}}_{\mathbf{c}}(\{\eta_j\}_j; \{\mathcal{C}_j^{(t)}\}_j) = \sum_{j=1}^{K} \sum_{x \in \mathcal{C}_j^{(t)}} \left[F_j^*(\eta_j) + k_j(x) + \langle t_j(x) - \eta_j, \nabla F_j^*(\eta_j) \rangle \right]. \tag{11.26}$$

When all $F_j^* = F^*$ are identical and the partition $\{\mathcal{C}_j^{(t)}\}_j$ corresponds to MAP assignment, $\mathring{\mathcal{L}}$ is exactly the objective function $\breve{\mathcal{L}}$ for the Bregman k-means [2]. Rewriting $\breve{\mathcal{L}}$ as an equivalent criterion to be minimized, it follows

$$\check{\mathcal{L}}(\{\eta_j\}_j) = \sum_{i=1}^{N} \min_{j=1}^{K} B_{F^*}(t(x_i) : \eta_j). \tag{11.27}$$

Bregman k-means++ [20, 21] provides initial centers $\{\eta_j^{(0)}\}_j$ which guarantee to find a clustering that is $\mathcal{O}(\log K)$-competitive to the optimal Bregman k-means clustering. The k-MLE++ algorithm amounts to use Bregman k-means++ on the dual log-normalizer F^* (see Algorithm 3).

Algorithm 3: k-MLE++

```
Input: A sample $\chi = \{x_1, ..., x_N\}$, $t$ the sufficient statistics and $F^*$ the dual log-normalizer
       of an exponential family, $K$ the number of clusters
Output: Initial mixture parameters $\{(w_j^{(0)}, \eta_j^{(0)}\}_j)$
$w_1^{(0)} = 1/K$;
Choose first seed $\eta_1^{(0)} = t(x_i)$ for $i$ uniformly random in $\{1, 2, \ldots, N\}$;
for $j = 2, ..., K$ do
    $w_j^{(0)} = 1/K$;
    // Compute relative contributions to $\check{\mathcal{L}}(\{\eta_j\}_j)$
    foreach $x_i \in \chi$ do $p_i = \frac{\min_{j'=1}^{j} B_{F^*}(t(x_i):\eta_{j'})}{\sum_{i'=1}^{N} \min_{j'=1}^{j} B_{F^*}(t(x_{i'}):\eta_{j'})}$;
    Choose $\eta_j^{(0)} \in \{t(x_1), ..., t(x_N)\}$ with probability $p_i$;
```

When K is unknown, same strategy can still be applied but a stopping criterion has to be set. Probability p_i in Algorithm 3 is a relative contribution of observation x_i through $t(x_i)$ to $\check{\mathcal{L}}(\{\eta_1, ..., \eta_K\})$ where K is the number of already selected centers. A high p_i indicates that x_i is relatively far from these centers, thus is atypic to the mixture $\{(w_1^{(0)}, \eta_1^{(0)}), ..., (w_K^{(0)}, \eta_K^{(0)}), \}$ for $w_j^{(0)} = w^{(0)}$ an arbitrary constant. When selecting a new center, p_i necessarily decreases in the next iteration. A good covering of χ is obtained when all p_i are lower than some threshold $\lambda \in [0, 1]$. Algorithm 4 describes the initialization named after DP-k-MLE++.

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

11.3.5 Initialization with DP-comp-k-MLE

Although k-MLE can be used with component-wise exponential families, previous initialization methods yield components of same exponential family. Component distribution may be chosen simultaneously to a center selection when additional

Algorithm 4: DP-k-MLE++

```
Input: A sample $\chi = \{x_1, ..., x_N\}$, $t$ the sufficient statistics and $F^*$ the dual log-normalizer
       of an exponential family, $\lambda \in [0, 1]$
Output: Initial mixture parameters $\{w_j^{(0)}, \eta_j^{(0)}\}_j$, $K$ the number of clusters
Choose first seed $\eta_1^{(0)} = t(x_i)$ for $i$ uniformly random in $\{1, 2, \dots, N\}$;
K=1;
repeat
    // Compute relative contributions to $\breve{\mathcal{L}}(\{\eta_1, ..., \eta_K\})$
    foreach $x_i \in \chi$ do $p_i = \frac{\min_{j=1}^{K} B_{F^*}(t(x_i):\eta_j)}{\sum_{i'=1}^{N} \min_{j=1}^{K} B_{F^*}(t(x_{i'}):\eta_{j'})}$;
    if $\exists\, p_i > \lambda$ then
        K = K+1;
        // Select next seed
        Choose $\eta_K^{(0)} \in \{t(x_1), ..., t(x_N)\}$ with probability $p_i$;
until all $p_i \leq \lambda$;
for $j = 1, ..., K$ do $w_j^{(0)} = 1/K$;
```

knowledge ξ_i about x_i is available (see Sect. 11.5 for an example). Given such a choice function H, Algorithm 5 called "DP-comp-k-MLE" describes this new flexible initialization method. DP-comp-k-MLE is clearly a generalization of DP-k-MLE++ when H always returns the same exponential family. However, in the general case, it remains to be proved whether a DP-comp-k-MLE clustering is $\mathcal{O}(\log K)$-competitive to the optimal k-MLE clustering (with equal weight). Without this difficult theoretical study, suffix "++" is carefully omitted in the name DP-comp-k-MLE.

To end up with this section, let us recall that all we need to know for using proposed algorithms is the MLE for the considered exponential family, whether it is available in a closed-form or not. In many exponential families, all details (canonical decomposition, $F, \nabla F, F^*, \nabla F^* = (\nabla F)^{-1}$) are already known [10]. The next section focuses on the case of the Wishart distribution.

11.4 Learning Mixtures of Wishart with k-MLE

This section recalls the definition of Wishart distribution and proposes a maximum likelihood estimator for its parameters. Some known facts such as the Kullback-Leibler divergence between two Wishart densities are recalled. Its use with the above algorithms is also discussed.

11.4.1 Wishart Distribution

The Wishart distribution [5] is the multidimensional version of the chi-square distribution and it characterizes empirical scatter matrix estimator for the multivariate gaussian distribution. Let $\mathbb{X}$ be a n-sample consisting in independent realizations of

Algorithm 5: DP-comp-k-MLE

```
Input: A sample $\chi = \{x_1, ..., x_N\}$ with extra knowledge $\xi = \{\xi_1, ..., \xi_N\}$, $H$ a choice
       function of an exponential family, $\lambda \in [0, 1]$
Output: Initial mixture parameters $\{(w_j^{(0)}, \eta_j^{(0)})\}_j$, $\{(t_j, F_j^*)\}_j$ sufficient statistics and dual
        log-normalizers of exponential families, $K$ the number of clusters
// Select first seed and exponential family
for $i$ uniformly random in $\{1, 2, \ldots, N\}$ do
    Obtain $t_1$, $F_1^*$ from $H(x_i, \xi_i)$;
    Select first seed $\eta_1^{(0)} = t_1(x_i)$;
K=1;
repeat
    foreach $x_i \in \chi$ do $p_i = \frac{\min_{j=1}^K B_{F_j^*}(t_j(x_i):\eta_j)}{\sum_{i'=1}^N \min_{j=1}^K B_{F_j^*}(t_j(x_{i'}):\eta_{j'})}$;
    if $\exists\ p_i > \lambda$ then
        K = K+1;
        // Select next seed and exponential family
        for $i$ with probability $p_i$ in $\{1, 2, \ldots, N\}$ do
            Obtain $t_K$, $F_K^*$ from $H(x_i, \xi_i)$;
            Select next seed $\eta_K^{(0)} = t_K(x_i)$;
until all $p_i \leq \lambda$;
for $j = 1, ..., K$ do $w_j^{(0)} = 1/K$;
```

a random gaussian vector with d dimensions, zero mean and covariance matrix S. Then scatter matrix $X = {}^t\mathbb{X}\mathbb{X}$ follows a central Wishart distribution with scale matrix S and degree of freedom n, denoted by $X \sim \mathcal{W}_d(n, S)$. Its density function is

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

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

Wishart distribution is an exponential family since

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

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

$$F(\theta_n, \theta_S) = \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|) + \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right). \tag{11.28}$$

Note that this decomposition is not unique (see another one in [22]). Refer to Appendix A.1 for detailed calculations.

11.4.2 MLE for Wishart Distribution

Let us recall (see Sect. 11.2.1) that the MLE is obtained by mapping $(\nabla F)^{-1}$ on the average of sufficient statistics. Finding $(\nabla F)^{-1}$ amounts to solve here the following system (see Eqs. 11.5 and 11.28):

$$\begin{cases} d\log(2) - \log|\theta_S| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right) = \eta_n, \\ -\left(\theta_n + \frac{(d+1)}{2}\right)\theta_S^{-1} = \eta_S. \end{cases} \tag{11.29}$$

with $\eta_n = \mathbb{E}[\log|X|]$ and $\eta_S = \mathbb{E}[-\frac{1}{2}X]$ the expectation parameters and Ψ_d the derivative of the $\log\Gamma_d$. Unfortunately, variables θ_n and θ_S are not separable so that no closed-form solution is known. Instead, as pointed out in [23], 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 $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$ of Wishart distribution $\mathcal{W}_d$ which are also exponential families. For the sake of simplicity, natural parameterizations and sufficient statistics of the decomposition in the general case are kept (see Appendices A.2 and A.3 for more details).

Distribution $\mathcal{W}_{d,\underline{n}}$ $(\underline{n} = 2\theta_{\underline{n}} + d + 1)$: $k_{\underline{n}}(X) = \frac{\underline{n}-d-1}{2}\log|X|$ and

$$F_{\underline{n}}(\theta_S) = \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|\theta_S| + \log\Gamma_d\left(\frac{\underline{n}}{2}\right). \tag{11.30}$$

Using classical results for matrix derivatives, (Eq. 11.5) can be easily solved:

$$-\frac{\underline{n}}{2}\hat{\theta}_S^{-1} = \frac{1}{N}\sum_{i=1}^{N} -\frac{1}{2}X_i \implies \hat{\theta}_S = N\underline{n}\left(\sum_{i=1}^{N} X_i\right)^{-1}. \tag{11.31}$$

Distribution $\mathcal{W}_{d,\underline{S}}$ $(\underline{S} = \theta_{\underline{S}}^{-1})$: $k_{\underline{S}}(X) = -\frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X)$ and

$$F_{\underline{S}}(\theta_n) = \left(\theta_n + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| + \log\Gamma_d\left(\theta_n + \frac{d+1}{2}\right) \tag{11.32}$$

Again, Eq. (11.5) can be numerically solved:

$$\hat{\theta}_n = \Psi_d^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\log|X_i| - \log\left|2\underline{S}\right|\right) - \frac{d+1}{2}, \qquad \hat{\theta}_n > -1 \tag{11.33}$$

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 method [24]). Let us mention that notation is simplified here since $\hat{\theta}_S$ and $\hat{\theta}_n$ should have been indexed by their corresponding family. Algorithm 6 summarizes the estimate $\hat{\theta}$ for parameters of the Wishart distribution. By precomputing $N\left(\sum_{i=1}^{N} X_i\right)^{-1}$ and $N^{-1}\sum_{i=1}^{N}\log|X_i|$, much computation time can be saved. The computation of the Ψ_d^{-1} remains an expensive part of the algorithm.

Let us now prove the convergence and the consistency of this method. Maximizing $\bar{l}$ amounts to minimize equivalently $E(\theta) = F(\theta) - \langle\frac{1}{N}\sum_{i=1}^{N} t(X_i), \theta\rangle$. The following properties are satisfied by E:

- The hessian $\nabla^2 E = \nabla^2 F$ of E is positive definite on Θ since F is convex.
- Its unique minimizer on Θ is the MLE $\hat{\theta} = \nabla F^*(\frac{1}{N}\sum_{i=1}^{N} t(X_i))$ whenever it exists (although F^* is not known for Wishart, and F is not separable).

Algorithm 6: MLE for the parameters of a Wishart distribution

```
Input: A sample $\chi = \{X_1, X_2, \dots, X_N\}$ of $\mathcal{S}^d_{++}$ with $N > 1$
Output: Estimate $\hat{\theta}$ is the terminal values of MLE sequences $\{\hat{\theta}_n^{(t)}\}$ and $\{\hat{\theta}_S^{(t)}\}$
// Initialization of the $\{\hat{\theta}_n^{(t)}\}$ sequence
$\hat{\theta}_n^{(0)} = 1; t = 0;$
repeat
    // Compute MLE in $\mathcal{W}_{d,\underline{n}}$ using Eq. 11.31
        $\hat{\theta}_S^{(t+1)} = N\underline{n}\left(\sum_{i=1}^{N} X_i\right)^{-1}$ with $\underline{n} = 2\hat{\theta}_n^{(t)} + d + 1$
    // Compute MLE in $\mathcal{W}_{d,\underline{S}}$ using Eq. 11.33
        $\hat{\theta}_n^{(t+1)} = \Psi_d^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\log|X_i| - \log\left|2\underline{S}\right|\right) - \frac{d+1}{2}$ with $\underline{S} = \left(\hat{\theta}_S^{(t+1)}\right)^{-1}$
    $t = t + 1;$
until convergence of the likelihood;
```

Therefore, Algorithm 6 is an instance of the group coordinate descent algorithm of Bezdek et al. (Theorem 2.2 in [25]) for $\theta = (\theta_n, \theta_S)$:

$$\hat{\theta}_S^{(t+1)} = \arg\max_{\theta_S} E(\hat{\theta}_n^{(t)}, \theta_S) \tag{11.34}$$

$$\hat{\theta}_n^{(t+1)} = \arg\max_{\theta_n} E(\theta_n, \hat{\theta}_S^{(t+1)}) \tag{11.35}$$

Resulting sequences $\{\hat{\theta}_n^{(t)}\}$ and $\{\hat{\theta}_S^{(t)}\}$ are shown to converge linearly to the coordinates of $\hat{\theta}$.

By looking carefully at the previous algorithms, let us remark that the initialization methods require to able to compute the divergence B_{F^*} between two elements η_1 and η_2 in the expectation space **H**. Whereas F^* is known for $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$, Eq. (11.9) gives a potential solution for $\mathcal{W}_d$ by considering B_F on natural parameters θ_2 and θ_1 in Θ. Searching the correspondence $\mathbf{H} \mapsto \Theta$ is analogous to compute the MLE for a single observation...

The previous MLE procedure does not converge with a single observation X_1. Bogdan and Bogdan [26] proved that MLE exists and is unique in an exponential family off the affine envelope of the N points $t(X_1), ..., t(X_N)$ is of dimension D, the order of this exponential family. Since the affine envelope of $t(X_1)$ is of dimension $d \times d$ (instead of D $= d \times d + 1$), the MLE does not exists and the likelihood function goes to infinity.[2] Unboundedness of likelihood function is well known problem that can be tackled by adding a penalty term to it [27]. A simpler solution is to take the MLE in family $\mathcal{W}_{d,\underline{n}}$ for some n (known or arbitrary fixed above $d - 1$) instead of $\mathcal{W}_d$.

11.4.3 Divergences for Wishart Distributions

For two Wishart distributions $\mathcal{W}_d^1 = \mathcal{W}_d(X; n_1, S_1)$ and $\mathcal{W}_d^2 = \mathcal{W}_d(X; n_2, S_2)$, the KL divergence is known [22] (even if F^* is unknown):

$$\begin{aligned}\mathrm{KL}(\mathcal{W}_d^1||\mathcal{W}_d^2) = &-\log\left(\frac{\Gamma_d\left(\frac{n_1}{2}\right)}{\Gamma_d\left(\frac{n_2}{2}\right)}\right) + \left(\frac{n_1 - n_2}{2}\right)\Psi_d\left(\frac{n_1}{2}\right)\\ &+ \frac{n_1}{2}\left(-\log\frac{|S_1|}{|S_2|} + \mathrm{tr}(S_2^{-1}S_1) - d\right)\end{aligned} \tag{11.36}$$

Looking the KL divergences of the two Wishart sub-families $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$ gives an interesting perspective to this formula. Applying Eqs. 11.9 and 11.8, it follows

$$\mathrm{KL}(\mathcal{W}_{d,\underline{n}}^1||\mathcal{W}_{d,\underline{n}}^2) = \frac{n}{2}\left(-\log\frac{|S_1|}{|S_2|} + \mathrm{tr}(S_2^{-1}S_1) - d\right) \tag{11.37}$$

$$\mathrm{KL}(\mathcal{W}_{d,\underline{S}}^1||\mathcal{W}_{d,\underline{S}}^2) = -\log\left(\frac{\Gamma_d\left(\frac{n_1}{2}\right)}{\Gamma_d\left(\frac{n_2}{2}\right)}\right) + \left(\frac{n_1 - n_2}{2}\right)\Psi_d\left(\frac{n_1}{2}\right) \tag{11.38}$$

Detailed calculations can be found in the Appendix. Notice that $\mathrm{KL}(\mathcal{W}_d^1||\mathcal{W}_d^2)$ is simply the sum of these two divergences

[2] Product $\hat{\theta}_n^{(t)}\hat{\theta}_S^{(t)}$ is constant through iterations.

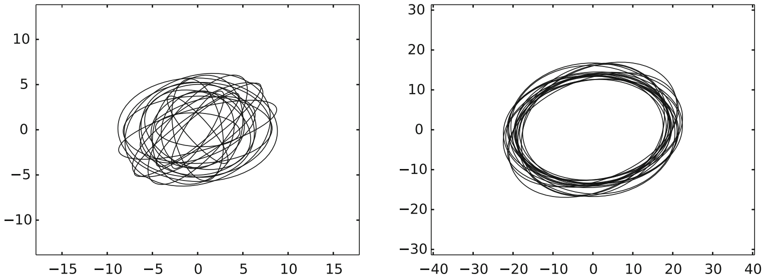

Fig. 11.1 20 random matrices from $\mathcal{W}_d(.; n, S)$ from $n = 5$ (*left*), $n = 50$ (*right*)

$$\mathrm{KL}(\mathcal{W}_d^1 || \mathcal{W}_d^2) = \mathrm{KL}(\mathcal{W}_{d,\underline{S_1}}^1 || \mathcal{W}_{d,\underline{S_1}}^2) + \mathrm{KL}(\mathcal{W}_{d,\underline{n_1}}^1 || \mathcal{W}_{d,\underline{n_1}}^2) \tag{11.39}$$

and that $\mathrm{KL}(\mathcal{W}_{d,\underline{S}}^1 || \mathcal{W}_{d,\underline{S}}^2)$ does not depend on $\underline{S}$.

Divergence $\mathrm{KL}(\mathcal{W}_{d,\underline{n}}^1 || \mathcal{W}_{d,\underline{n}}^2)$, commonly used as a dissimilarity measure between covariance matrices, is sometimes referred as the log-Det divergence due to the form of $\varphi(S) = F_{\underline{n}}(S) \propto \log |S|$ (see Eq. 11.30). However, the dependency on term $\underline{n}$ should be neglected only when the two empirical covariance matrices comes from samples of the same size. In this case, log-Det divergence between two covariance matrices is the KL divergence in the sub-family $\mathcal{W}_{d,\underline{n}}$.

11.4.4 Toy Examples

In this part, some simple simulations are given for $d = 2$. Since the observations are positive semi-definite matrices, it is possible to visualize them with ellipses parametrized by their eigen decompositions. For example, Fig. 11.1 shows 20 matrices generated from $\mathcal{W}_d(.; n, S)$ for $n = 5$ and for $n = 50$ with S having eigenvalues $\{2, 1\}$. This visualization highlights the difficulty for the estimation of the parameters (even for d small) when n is small.

Then, a dataset of 60 matrices is generated from a three components mixture with parameters $\mathcal{W}_d(.; 10, S_1)$, $\mathcal{W}_d(.; 20, S_2)$, $\mathcal{W}_d(.; 30, S_3)$ and equal weights $w_1 = w_2 = w_3 = 1/3$. The respective eigenvalues for S_1, S_2, S_3 are in turn $\{2, 1\}, \{2, 0.5\}, \{1, 1\}$. Figure 11.2 illustrates this dataset. To study the influence of a good initialization for k-MLE, the Normalized Mutual Information (NMI) [28] is computed between the final partition and the ground-truth partition for different initializations. This value between 0 and 1 is higher when the two partitions are more similar. Following table gives average and standard deviation of NMI over 30 runs:

	Rand. Init/Lloyd	Rand. Init/Hartigan	k-MLE++/Hartigan
NMI	0.229 ± 0.279	0.243 ± 0.276	0.67 ± 0.083

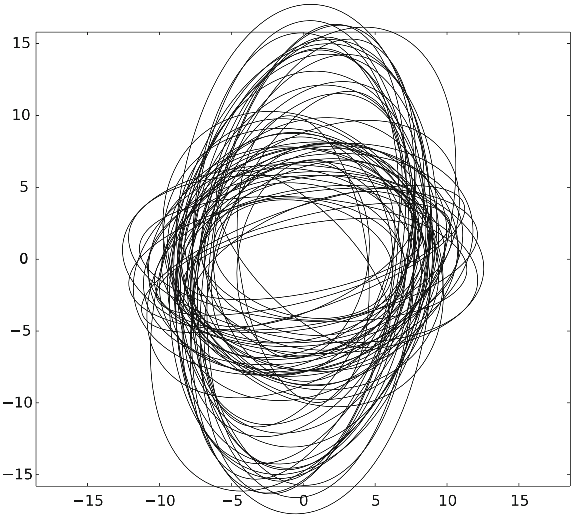

Fig. 11.2 Dataset: 60 matrices generated from a three components mixture of $\mathcal{W}_2$

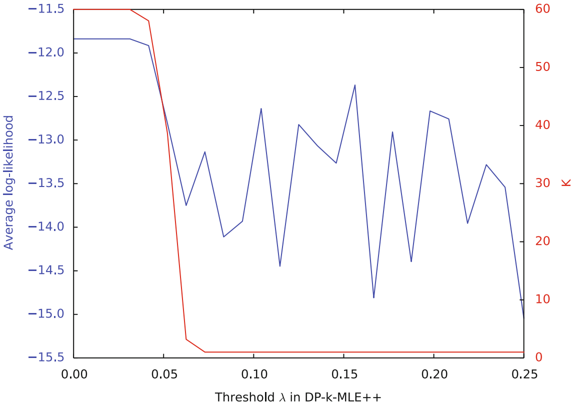

Fig. 11.3 DP-k-MLE++: Influence of λ on K and on the average log-likelihood

From this small experiment, we can easily verify the importance of a good initialization. Also, the partitions having the highest NMI are reported in Fig. 11.4 for each method. Let us mention that Hartigan method gives almost always a better partition than the Lloyd's one for the same initial mixture.

A last simulation indicates that the initialization with DP-k-MLE++ is very sensible to its parameter λ. Again with the same set of matrices, Fig. 11.3 shows how the number of generated clusters K and the average log-likelihood evolve with λ. Not surprisingly, both quantities decrease when λ increases.

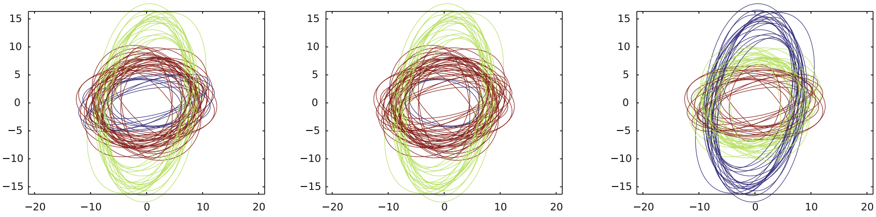

Fig. 11.4 Best partitions with Rand. Init/Lloyd (*left*), Rand. Init/Hartigan (*middle*), k-MLE++ Hartigan (*right*)

11.5 Application to Motion Retrieval

In this section, a potential application to motion retrieval is proposed following our previous work [23]. Raw motion-captured movement can be identifiable to a $n_i \times d$ matrix $\mathbb{X}_i$ where each row corresponds to captured locations of a set of sensors.

11.5.1 Movement Representation

When the aim is to provide a taxonomy of a set of movements, it is difficult to compare varying-size matrices. Cross-product matrices $X_i = {}^t\mathbb{X}_i\mathbb{X}_i$ is a possible descriptor[3] of $\mathbb{X}_i$. Denoting N the number of movements, set $\{X_1, ..., X_N\}$ of $d \times d$ matrix is exactly the input of k-MLE. Note that d can easily be of high dimension when the number of sensors is large.

The simplest way to initialize k-MLE in this setting is to apply DP-k-MLE++ for $\mathcal{W}_d$. But when n_i are known, it is better not to estimate them. In this case, DP-comp-k-MLE is appropriate for a function H selecting $\mathcal{W}_{d,\underline{n_i}}$ given $\xi_i = n_i$. When learning algorithm is fast enough, it is common practice to restart it for different initializations and to keep the best output (mixture parameters).

To enrich the description of a single movement, it is possible to define a mixture m_i per movement $\mathbb{X}_i$. For example, several subsets of successive observations with different sizes can be extracted and their cross-product matrices used as inputs for k-MLE (and DP-comp-k-MLE). Mixture m_i can be viewed as a sparse representation of local dynamics of $\mathbb{X}_i$ through their local second-order moments.

While these two representations are of different kind, it is possible to encompass both in a common framework for $\mathbb{X}_i$ described by a mixture of a single component $\{(w_{i,1} = 1, \eta_{i,1} = t(X_i))\}$. Algorithm k-MLE applied on such input for all movements (i.e. $\{t(X_i)\}_i$) provides then another set of mixture parameters $\{(\hat{w}_j, \hat{\eta}_j)\}_j$. Note that the general treatment of arbitrary mixtures of mixtures of Wishart is not claimed to be addressed here.

[3] For translation invariance, $\mathbb{X}_i$ are column centered before.

11.5.2 Querying with Cauchy-Schwartz Divergence

Let us consider a movement $\mathbb{X}$ (a $n \times d$ matrix) and its mixture representation m. Without loss of generality, let us denote $\{(w_j, \theta_j)\}_{j=1..K}$ the mixture parameters for m. The problem of comparing two movements amounts to compute a appropriate dissimilarity between m and another mixture m' of such a kind with parameters $\{(w'_j, \theta'_j)\}_{j=1..K'}$.

When both mixtures have a single component (K = K' = 1), an immediate solution is to consider the Kullback-Leibler divergence KL$(m : m)$ for two members of the same exponential family. Since it is the Bregman divergence on the swapped natural parameters $B_F(\theta' : \theta)$, a closed form is always available from Eq. (11.7). It is important to mention that this formula holds for θ and θ' viewed as parameters for $\mathcal{W}_d$ even if they are estimated in sub-families $W_{d,n}$ and $W_{d,n'}$.

For general mixtures of the same exponential family ($K > 1$ or $K' > 1$), KL divergence admits no more a closed form and has to be approximate with numerical methods. Recently, other divergences such as the Cauchy-Schwartz divergence (CS) [29] were shown to be available in a closed form:

$$\mathrm{CS}(m : m') = -\log \frac{\int m(x)m'(x)\mathrm{d}x}{\sqrt{\int m(x)^2 \mathrm{d}x \int m'(x)^2 \mathrm{d}x}}. \tag{11.40}$$

Within the same exponential family p_F, the integral of the product of mixtures is

$$\int m(x)m'(x)\mathrm{d}x = \sum_{j=1}^{K}\sum_{j'=1}^{K'} w_j w'_{j'} \int p_F(x;\theta_j)p_F(x;\theta'_{j'})\mathrm{d}x. \tag{11.41}$$

When carrier measure $k(X) = 0$, as it is for $\mathcal{W}_d$ but not for $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$, the integral can be further expanded as

$$\begin{aligned}\int p_F(x;\theta_j)p_F(x;\theta'_{j'})\mathrm{d}X &= \int e^{\langle\theta_j, t(X)\rangle - F(\theta_j)} e^{\langle\theta'_j, t(X)\rangle - F(\theta'_{j'})}\mathrm{d}X \\ &= \int e^{\langle\theta_j+\theta'_{j'}, t(X)\rangle - F(\theta_j) - F(\theta'_{j'})}\mathrm{d}X \\ &= e^{F(\theta_j+\theta'_{j'}) - F(\theta_j) - F(\theta'_{j'})} \underbrace{\int e^{\langle\theta_j+\theta'_{j'}, t(X)\rangle - F(\theta_j+\theta'_{j'})}\mathrm{d}X}_{=1}.\end{aligned}$$

Note that $\theta_j + \theta'_{j'}$ must be in the natural parameter space Θ to ensure that $F(\theta_j + \theta'_{j'})$ is finite. An equivalent condition is that Θ is a *convex cone*.

When $p_F = \mathcal{W}_d$, space $\Theta =]-1; +\infty[\times \mathcal{S}^p_{++}$ is not a convex cone since $\theta_{n_j} + \theta'_{n_{j'}} < -1$ for n_j and $n'_{j'}$ smaller than $d+1$. Practically, this constraint is tested for each parameter pairs before going on with the computation the CS divergence. A possible fix, not developed here, would be to constraint n to be greater than $d + 1$ (or equivalently $\theta_n > 0$). Such a constraint amounts to take a convex subset $]0; +\infty[\times \mathcal{S}^p_{++}$ of Θ. Denoting $\Delta(\theta_j, \theta'_{j'}) = F(\theta_j + \theta'_{j'}) - F(\theta_j) - F(\theta'_{j'})$, the CS divergence is also

$$\begin{aligned} \mathrm{CS}(m : m') &= \frac{1}{2} \log \sum_{j=1}^{K} \sum_{j'=1}^{K} \left[w_j w_{j'} \exp^{\Delta(\theta_j, \theta_{j'})} \right] && (within\ m) \\ &+ \frac{1}{2} \log \sum_{j=1}^{K'} \sum_{j'=1}^{K'} \left[w'_j w'_{j'} \exp^{\Delta(\theta'_j, \theta'_{j'})} \right] && (within\ m') \\ &- \log \sum_{j=1}^{K} \sum_{j'=1}^{K'} \left[w_j w'_{j'} \exp^{\Delta(\theta_j, \theta'_{j'})} \right] && (between\ m\ and\ m') \end{aligned} \tag{11.42}$$

Note that CS divergence is symmetric since $\Delta(\theta_j, \theta'_{j'})$ is. A numeric value of $\Delta(\theta_j, \theta'_{j'})$ can be computed for $\mathcal{W}_d$ from Eq. 11.28 (see Eq. 11.45 or 11.46 in the Appendix).

11.5.3 Summary of Proposed Motion Retrieval System

To conclude this section, let us recall the elements of our proposal for a motion retrieval system. Movement is represented by a Wishart mixture model learned by k-MLE initialized by DP-k-MLE++ or DP-comp-k-MLE. In the case of a mixture of a component, a simple application of the MLE for $\mathcal{W}_{d,\underline{n}}$ is sufficient. Although a Wishart distribution appears inadequate model for the scatter matrix X of a movement, it has been shown that this crude assumption provides a good classification rates on a real data set [23]. Learning representations of the movements may be performed offline since it is computational demanding. Using CS divergence as dissimilarity, we can then extract a taxonomy of movements with any spectral clustering algorithm. For a query movement, its representation by a mixture has to be computed first. Then it is possible to search the database for the most similar movements according to the CS divergence or to predict its type by a majority vote among them. More details of the implementation and results for the real dataset will be in a forthcoming technical report.

11.6 Conclusions and Perspectives

Hartigan's swap clustering method for k-MLE was studied for the general case of an exponential family. Unlike for k-means, this method does not guarantee to avoid empty clusters but achieves generally better performance than the Lloyd's heuristic.

Two methods DP-k-MLE and DP-comp-k-MLE are proposed to initialize k-MLE automatically by setting the number of clusters. While the former shares the good properties of k-MLE, the latter selects the component distributions given some extra knowledge. A small experiment indicates these methods appear to be quite sensible to their only parameter.

We recalled the definition and some properties of the Wishart distribution $\mathcal{W}_d$, especially its canonical decomposition as a member of an exponential family.By fixing either one of its two parameters n and S, two other (nested) exponential (sub-) families $\mathcal{W}_{d,\underline{n}}$ and $\mathcal{W}_{d,\underline{S}}$ may be defined. From their respective MLEs, it is possible to define an iterative process which provably converges to the MLE for $\mathcal{W}_d$. For a single observation, the MLE does not exist.Then a crude solution is to replace the MLE in $\mathcal{W}_d$ by the MLE in one of the two sub-families.

The MLE is an example of a point estimator among many others (e.g. method of moments, minimax estimators, Bayesian point estimators). This suggests as future work many other learning algorithms such as k-MoM, k-Minimax [30], k-MAP following the same algorithmic scheme as k-MLE.

Finally, an application to the retrieval motion-captured motions is proposed.Each motion is described by a Wishart mixture model and the Cauchy-Schwarz divergence is used as a dissimilarity measure between two mixture models.As the CS divergence is always available in closed-form, such divergence is fast to compute compared to stochastic integration estimation schemes.This divergence can be used in spectral clustering methods and for visualization of a set of motions in an Euclidean embedding.

Another perspective is the connection between the closed-form divergences between mixtures and kernels based on divergences [31]: The CS divergence looks similar to the Normalized Correlation Kernel [32].This could lead to a broader class of methods (e.g., SVM) using these divergences.

Appendix A

This Appendix details some calculations for distributions $\mathcal{W}_d$, $\mathcal{W}_{d,\underline{n}}$, $\mathcal{W}_{d,\underline{S}}$.

A.1 Wishart Distribution $\mathcal{W}_d$

$$\begin{aligned}\mathcal{W}_d(X; n, S) &= \frac{|X|^{\frac{n-d-1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(S^{-1}X)\}}{2^{\frac{nd}{2}} |S|^{\frac{n}{2}} \Gamma_d(\frac{n}{2})} \\ &= \exp\left\{\frac{n-d-1}{2} \log|X| - \frac{1}{2}\mathrm{tr}(S^{-1}X) - \frac{nd}{2}\log(2) - \frac{n}{2}\log|S| - \log\Gamma_d\left(\frac{n}{2}\right)\right\}\end{aligned}$$

Letting $(\theta_n, \theta_S) = (\frac{n-d-1}{2}, S^{-1}) \longleftrightarrow (n, S) = (2\theta_n + d + 1, \theta_S^{-1})$

$$\begin{aligned}
\mathcal{W}_d(X;\theta_n,\theta_S) =& \exp\left\{\frac{2\theta_n+d+1-d-1}{2}\log|X| - \frac{1}{2}\mathrm{tr}(\theta_S X) - \frac{(2\theta_n+d+1)d}{2}\log(2)\right.\\
&\left. - \frac{(2\theta_n+d+1)}{2}\log|\theta_S^{-1}| - \log\Gamma_d\left(\frac{2\theta_n+d+1}{2}\right)\right\}\\
=& \exp\left\{\theta_n\log|X| - \frac{1}{2}\mathrm{tr}(\theta_S X) - \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|)\right.\\
&\left. - \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right)\right\}\\
=& \exp\left\{< \theta_n, \log|X| >_{\mathbb{R}} + < \theta_S, -\frac{1}{2}X >_{HS} - F(\Theta)\right\}\\
&\text{with } F(\Theta) = \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|) + \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right)\\
=& \exp\{< \Theta, t(X) > - F(\Theta) + k(X)\}\\
&\text{with} t(X) = (\log|X|, -\frac{1}{2}X) \text{ and } k(X) = 0
\end{aligned}$$

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

$$\frac{\partial F}{\partial\theta_n}(\theta_n,\theta_S) = d\log(2) - \log|\theta_S| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right) \tag{11.43}$$

where Ψ_d is the multivariate Digamma function (or multivariate polygamma of order 0).

$$\frac{\partial F}{\partial\theta_S}(\theta_n,\theta_S) = -\left(\theta_n + \frac{(d+1)}{2}\right)\theta_S^{-1} \tag{11.44}$$

Dissimilarity $\Delta(\theta,\theta')$ between natural parameters $\theta = (\theta_n, \theta_S)$ and $\theta' = (\theta'_n, \theta'_S)$ is

$$\begin{aligned}
\Delta(\theta,\theta') =& F(\theta+\theta') - (F(\theta)+F(\theta')) = \left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right)\left(d\log(2) - \log|\theta_S + \theta'_S|\right)\\
&- \left(\theta_n + \frac{(d+1)}{2}\right)(d\log(2) - \log|\theta_S|) - \left(\theta'_n + \frac{(d+1)}{2}\right)\left(d\log(2) - \log|\theta'_S|\right)\\
&+ \log\Gamma_d\left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right) - \log\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right) - \log\Gamma_d\left(\theta'_n + \frac{(d+1)}{2}\right)\\
=& -\frac{(d+1)}{2}d\log(2) + \left(\theta_n + \frac{(d+1)}{2}\right)\log|\theta_S| + \left(\theta'_n + \frac{(d+1)}{2}\right)\log|\theta'_S|\\
&- \left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right)\log|\theta_S + \theta'_S| + \log\left(\frac{\Gamma_d\left(\theta_n + \theta'_n + \frac{(d+1)}{2}\right)}{\Gamma_d\left(\theta_n + \frac{(d+1)}{2}\right)\Gamma_d\left(\theta'_n + \frac{(d+1)}{2}\right)}\right)
\end{aligned} \tag{11.45}$$

Remark $\Delta(\theta, \theta) \neq 0$. Same quantity with source parameters $\lambda = (n, S)$ and $\lambda' = (n', S')$ is

$$\begin{aligned}\Delta(\lambda, \lambda') = &- \frac{(d+1)}{2} d\log(2) - \frac{n}{2}\log|S| - \frac{n'}{2}\log|S'| - \frac{n+n'-d-1}{2}\log|S^{-1} \\ &+ S'^{-1}| + \log\left(\frac{\Gamma_d\left(\frac{n+n'-d-1}{2}\right)}{\Gamma_d\left(\frac{n}{2}\right)\Gamma_d\left(\frac{n'}{2}\right)}\right) \qquad (11.46)\end{aligned}$$

A.2 Distribution $\mathcal{W}_{d,\underline{n}}$

$$\begin{aligned}\mathcal{W}_d(X; \underline{n}, S) &= \frac{|X|^{\frac{\underline{n}-d-1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(S^{-1}X)\}}{2^{\frac{\underline{n}d}{2}}|S|^{\frac{\underline{n}}{2}}\Gamma_d(\frac{\underline{n}}{2})} \\ &= \exp\left\{\frac{\underline{n}-d-1}{2}\log|X| - \frac{1}{2}\mathrm{tr}(S^{-1}X) - \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|S| - \log\Gamma_d\left(\frac{\underline{n}}{2}\right)\right\}\end{aligned}$$

Letting $\theta_S = S^{-1}$,

$$\begin{aligned}\mathcal{W}_d(X; \underline{n}, \theta_S) &= \exp\left\{-\frac{1}{2}\mathrm{tr}(\theta_S X) + \frac{\underline{n}-d-1}{2}\log|X| - \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|\theta_S^{-1}| - \log\Gamma_d\left(\frac{\underline{n}}{2}\right)\right\} \\ &= \exp\left\{< \theta_S, -\frac{1}{2}X >_{HS} + k(X) - F_{\underline{n}}(\theta_S)\right\} \\ &\quad \text{with } F_{\underline{n}}(\theta_S) = \frac{\underline{n}d}{2}\log(2) - \frac{\underline{n}}{2}\log|\theta_S| + \log\Gamma_d\left(\frac{\underline{n}}{2}\right) \\ &\quad \text{with } k_{\underline{n}}(X) = \frac{\underline{n}-d-1}{2}\log|X|\end{aligned}$$

Using the rule $\frac{\partial log|X|}{\partial X} =^t (X^{-1})$ [33] and the symmetry of θ_S, we get

$$\nabla_{\theta_S} F_{\underline{n}}(\theta_S) = -\frac{\underline{n}}{2}\theta_S^{-1}$$

The correspondence between natural parameter θ_S and expectation parameter η_S is

$$\eta_S = \nabla_{\theta_S} F_{\underline{n}}(\theta_S) = -\frac{\underline{n}}{2}\theta_S^{-1} \longleftrightarrow \theta_S = \nabla_{\eta_S} F_{\underline{n}}^*(\eta_S) = (\nabla_{\theta_S} F_{\underline{n}})^{-1}(\eta_S) = -\frac{\underline{n}}{2}\eta_S^{-1}$$

Finally, we obtain the MLE for θ_S in this sub family:

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

Same formulation with source parameter S:

$$\hat{S} = \hat{\theta}_S^{-1} = \left(\underline{n}N\left(\sum_{i=1}^{N} X_i\right)^{-1}\right)^{-1} = \frac{\sum_{i=1}^{N} X_i}{\underline{n}N}$$

Dual log-normalizer $F^*_{\underline{n}}$ for $\mathcal{W}_{d,\underline{n}}$ is

$$\begin{aligned}
F^*_{\underline{n}}(\eta_S) &= \langle(\nabla F_{\underline{n}})^{-1}(\eta_S), \eta_S\rangle - F_{\underline{n}}((\nabla F_{\underline{n}})^{-1}(\eta_S)) \\
&= \langle -\frac{n}{2}\eta_S^{-1}, \eta_S\rangle - F_{\underline{n}}(-\frac{n}{2}\eta_S^{-1}) \\
&= -\frac{n}{2}\mathrm{tr}(\eta_S^{-1}\eta_S) - \frac{nd}{2}\log(2) + \frac{n}{2}\log\left[(\frac{n}{2})^d| - \eta_S^{-1}|\right] - \log\Gamma_d\left(\frac{n}{2}\right) \\
&= -\frac{nd}{2}(1 + \log(2) - \log\underline{n} + \log 2) + \frac{n}{2}\log| - \eta_S^{-1}| - \log\Gamma_d\left(\frac{n}{2}\right) \\
&= \frac{nd}{2}\log\left(\frac{n}{4e}\right) + \frac{n}{2}\log| - \eta_S^{-1}| - \log\Gamma_d\left(\frac{n}{2}\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}(\mathcal{W}^1_{d,\underline{n}}||\mathcal{W}^2_{d,\underline{n}}) &= B_{F_{\underline{n}}}(\theta_{S_2} : \theta_{S_1}) \\
&= F_{\underline{n}}(\theta_{S_2}) - F_{\underline{n}}(\theta_{S_1}) - <\theta_{S_2} - \theta_{S_1}, \nabla_{\theta_S} F_{\underline{n}}(\theta_{S_1})> \\
&= \frac{n}{2}\left(\log|\theta_{S_1}| - \log|\theta_{S_2}|\right) + \frac{n}{2}\mathrm{tr}((\theta_{S_2} - \theta_{S_1})\theta_{S_1}^{-1}) \\
&= \frac{n}{2}\left(\log\frac{|\theta_{S_1}|}{|\theta_{S_2}|} + \mathrm{tr}(\theta_{S_2}\theta_{S_1}^{-1}) - d\right) \\
&= \frac{n}{2}\left(-\log\frac{|\theta_{S_2}|}{|\theta_{S_1}|} + \mathrm{tr}(\theta_{S_2}\theta_{S_1}^{-1}) - d\right)
\end{aligned}$$

also with source parameter

$$\mathrm{KL}(\mathcal{W}^1_{d,\underline{n}}||\mathcal{W}^2_{d,\underline{n}}) = \frac{n}{2}\left(-\log\frac{|S_1|}{|S_2|} + \mathrm{tr}(S_2^{-1}S_1) - d\right)$$

Let's remark that KL divergence depends now on $\underline{n}$.

$$\begin{aligned}
B_{F^*_{\underline{n}}}(\eta_{S_1} : \eta_{S_2}) &= F^*_{\underline{n}}(\eta_{S_1}) - F^*_{\underline{n}}(\eta_{S_2}) - <\eta_{S_1} - \eta_{S_2}, \nabla F^*_{\underline{n}}(\eta_{S_2})>_{HS} \\
&= \frac{n}{2}\left(\log| - \eta_{S_1}^{-1}| - \log| - \eta_{S_2}^{-1}|\right) - <\eta_{S_1} - \eta_{S_2}, -\frac{n}{2}\eta_{S_2}^{-1}>_{HS} \\
&= \frac{n}{2}\left(\log\frac{| - \eta_{S_1}^{-1}|}{| - \eta_{S_2}^{-1}|} + \mathrm{tr}(\eta_{S_1}\eta_{S_2}^{-1}) - d\right)
\end{aligned}$$

A.3 Distribution $\mathcal{W}_{d,\underline{S}}$

For fixed $\underline{S}$, the p.d.f of $\mathcal{W}_{d,\underline{S}}$ can be rewritten[4] as

$$\begin{aligned}\mathcal{W}_d(X;n,\underline{S}) &= \frac{|X|^{\frac{n-d-1}{2}}\exp\{-\frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X)\}}{|2\underline{S}|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})}\\ &= \exp\left\{\frac{n-d-1}{2}\log|X| - \frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X) - \frac{n}{2}\log|2\underline{S}| - \log\Gamma_d\left(\frac{n}{2}\right)\right\}\end{aligned}$$

Letting $\theta_n = \frac{n-d-1}{2}$ $(n = 2\theta_n + d + 1)$

$$\begin{aligned}\mathcal{W}_d(X;\theta_n,\underline{S}) &= \exp\left\{\theta_n\log|X| - \frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X) - \left(\theta_n + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| - \log\Gamma_d\left(\theta_n + \frac{d+1}{2}\right)\right\}\\ &= \exp\left\{<\theta_n, \log|X|> + k_{\underline{S}}(X) - F_{\underline{S}}(\theta_n)\right\}\\ &\quad \text{with } F_{\underline{S}}(\theta_n) = \left(\theta_n + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| + \log\Gamma_d\left(\theta_n + \frac{d+1}{2}\right)\\ &\quad \text{with } k_{\underline{S}}(X) = -\frac{1}{2}\mathrm{tr}(\underline{S}^{-1}X)\end{aligned}$$

The correspondence between natural parameter θ_n and expectation parameter η_n is

$$\eta_n = \nabla_{\theta_n}F_{\underline{S}}(\theta_n) = \log\left|2\underline{S}\right| + \Psi_d\left(\theta_n + \frac{(d+1)}{2}\right)$$

$$\begin{aligned}\Leftrightarrow &\quad \Psi_d\left(\theta_n + \tfrac{(d+1)}{2}\right) = \eta_n - \log\left|2\underline{S}\right|\\ \Leftrightarrow &\quad \theta_n + \tfrac{(d+1)}{2} = \Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\\ \Leftrightarrow &\ \theta_n = \Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right) - \tfrac{(d+1)}{2} = (\nabla F_{\underline{S}})^{-1}(\eta_n) = \nabla F^*_{\underline{S}}(\eta_n)\end{aligned}$$

Finally, we obtain the MLE for θ_n in this sub family:

$$\hat{\theta}_n = \Psi_d^{-1}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\log|X|\right] - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}$$

Same formulation with source parameter n:

[4] Since $|2S| = 2^d|S|$, we have $2^{\frac{nd}{2}}|S|^{\frac{n}{2}}$ that is equivalent to $|2S|^{\frac{n}{2}}$.

$$\frac{\hat{n}-d-1}{2} = \Psi_d^{-1}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\log|X|\right] - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}$$

$$\hat{n} = 2\Psi_d^{-1}\left(\left[\frac{1}{N}\sum_{i=1}^{N}\log|X|\right] - \log\left|2\underline{S}\right|\right)$$

Dual log-normalizer $F^*_{\underline{S}}$ for $\mathcal{W}_{d,\underline{S}}$ is

$$\begin{aligned}
F^*_{\underline{S}}(\eta_n) &= \langle(\nabla F_{\underline{S}})^{-1}(\eta_n), \eta_n\rangle - F_{\underline{S}}((\nabla F_{\underline{S}})^{-1}(\eta_n))\\
&= \langle \Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}, \eta_n\rangle\\
&\quad - \Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\log\left|2\underline{S}\right| - \log\Gamma_d\left(\Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\right)\\
&= \Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\left(\eta_n - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}\eta_n - \log\Gamma_d\left(\Psi_d^{-1}\left(\eta_n - \log\left|2\underline{S}\right|\right)\right)
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}(\mathcal{W}^1_{d,\underline{S}}||\mathcal{W}^2_{d,\underline{S}}) &= B_{F_{\underline{S}}}(\theta_{n_2}:\theta_{n_1}) = F_{\underline{S}}(\theta_{n_2}) - F_{\underline{S}}(\theta_{n_1}) - \langle\theta_{n_2}-\theta_{n_1}, \nabla F_{\underline{S}}(\theta_{n_1})\rangle\\
&= \left(\theta_{n_2} + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| + \log\Gamma_d\left(\theta_{n_2} + \frac{d+1}{2}\right)\\
&\quad - \left(\theta_{n_1} + \frac{d+1}{2}\right)\log\left|2\underline{S}\right| - \log\Gamma_d\left(\theta_{n_1} + \frac{d+1}{2}\right)\\
&\quad - \langle\theta_{n_2}-\theta_{n_1}, \log\left|2\underline{S}\right| + \Psi_d\left(\theta_{n_1} + \frac{(d+1)}{2}\right)\rangle
\end{aligned}$$

$$\begin{aligned}
\mathrm{KL}(\mathcal{W}^1_{d,\underline{S}}||\mathcal{W}^2_{d,\underline{S}}) &= \log\frac{\Gamma_d\left(\theta_{n_2} + \frac{d+1}{2}\right)}{\Gamma_d\left(\theta_{n_1} + \frac{d+1}{2}\right)} - (\theta_{n_2}-\theta_{n_1})\Psi_d\left(\theta_{n_1} + \frac{(d+1)}{2}\right)\\
&= -\log\frac{\Gamma_d\left(\theta_{n_1} + \frac{d+1}{2}\right)}{\Gamma_d\left(\theta_{n_2} + \frac{d+1}{2}\right)} + (\theta_{n_1}-\theta_{n_2})\Psi_d\left(\theta_{n_1} + \frac{(d+1)}{2}\right)
\end{aligned}$$

also with source parameter

$$\mathrm{KL}(\mathcal{W}^1_{d,\underline{S}}||\mathcal{W}^2_{d,\underline{S}}) = -\log\left(\frac{\Gamma_d\left(\frac{n_1}{2}\right)}{\Gamma_d\left(\frac{n_2}{2}\right)}\right) + \left(\frac{n_1-n_2}{2}\right)\Psi_d\left(\frac{n_1}{2}\right)$$

Let us remark that this quantity does not depend on $\underline{S}$.

$$\begin{aligned}
B_{F^*_{\underline{S}}}(\eta_{n_1}:\eta_{n_2}) &= F^*_{\underline{S}}(\eta_{n_1}) - F^*_{\underline{S}}(\eta_{n_2}) - <\eta_{n_1}-\eta_{n_2}, \nabla F^*_{\underline{S}}(\eta_{n_2})>_{HS}\\
&= \Psi_d^{-1}\left(\eta_{n_1} - \log\left|2\underline{S}\right|\right)\left(\eta_{n_1} - \log\left|2\underline{S}\right|\right) - \frac{(d+1)}{2}\eta_{n_1}\\
&\quad - \log\Gamma_d\left(\Psi_d^{-1}\left(\eta_{n_1} - \log\left|2\underline{S}\right|\right)\right)\\
&\quad - \Psi_d^{-1}\left(\eta_{n_2} - \log\left|2\underline{S}\right|\right)\left(\eta_{n_2} - \log\left|2\underline{S}\right|\right) + \frac{(d+1)}{2}\eta_{n_2}
\end{aligned}$$

$$
\begin{aligned}
&+ \log \Gamma_d \left(\Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) \right) \\
&- \langle \eta_{n_1} - \eta_{n_2}, \Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) - \frac{(d+1)}{2} \rangle_{HS}
\end{aligned}
$$

$$
\begin{aligned}
B_{F^*_{\underline{S}}}(\eta_{n_1} : \eta_{n_2}) = \log \frac{\Gamma_d \left(\Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) \right)}{\Gamma_d \left(\Psi_d^{-1} \left(\eta_{n_1} - \log \left| 2\underline{S} \right| \right) \right)} \\
- \left[\Psi_d^{-1} \left(\eta_{n_2} - \log \left| 2\underline{S} \right| \right) - \Psi_d^{-1} \left(\eta_{n_1} - \log \left| 2\underline{S} \right| \right) \right] \left(\eta_{n_1} - \log \left| 2\underline{S} \right| \right)
\end{aligned}
$$

References

1. McLachlan, G.J., Krishnan, T.: The EM Algorithm and Extensions, 2nd edn. Wiley Series in Probability and Statistics. Wiley-Interscience, New York (2008)
2. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. J. Mach. Learn. Res. **6**, 1705–1749 (2005)
3. Nielsen, F.: *k*-MLE: a fast algorithm for learning statistical mixture models. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 869–872 (2012). Long version as arXiv:1203.5181
4. Jain, A.K.: Data clustering: 50 years beyond *K*-means. Pattern Recogn. Lett. **31**, 651–666 (2010)
5. Wishart, J.: The generalised product moment distribution in samples from a Normal multivariate population. Biometrika **20**(1/2), 32–52 (1928)
6. Tsai, M.-T.: Maximum likelihood estimation of Wishart mean matrices under Lwner order restrictions. J. Multivar. Anal. **98**(5), 932–944 (2007)
7. Formont, P., Pascal, T., Vasile, G., Ovarlez, J.-P., Ferro-Famil, L.: Statistical classification for heterogeneous polarimetric SAR images. IEEE J. Sel. Top. Sign. Proces. **5**(3), 567–576 (2011)
8. Jian, B., Vemuri, B.: Multi-fiber reconstruction from diffusion MRI using mixture of wisharts and sparse deconvolution. In: Information Processing in Medical Imaging, pp. 384–395, Springer, Berlin (2007)
9. Cherian, A., Morellas, V., Papanikolopoulos, N., Bedros, S.: Dirichlet process mixture models on symmetric positive definite matrices for appearance clustering in video surveillance applications. In: Computer Vision and Pattern Recognition (CVPR), pp. 3417–3424 (2011)
10. Nielsen, F., Garcia, V.: Statistical exponential families: a digest with flash cards. http://arxiv.org/abs/0911.4863.. Accessed Nov 2009
11. Rockafellar, R.T.: Convex Analysis, vol. 28. Princeton University Press, Princeton (1997)
12. Wainwright, M.J., Jordan, M.J.: Graphical models, exponential families, and variational inference. Found. Trends Mach. Learn. **1**(1–2), 1–305 (2008)
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. (Methodological). **39** 1–38 (1977)
14. Celeux, G., Govaert, G.: A classification EM algorithm for clustering and two stochastic versions. Comput. Stat. Data Anal. **14**(3), 315–332 (1992)
15. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A *k*-means clustering algorithm. J. Roy. Stat. Soc. C (Applied Statistics). **28**(1), 100–108 (1979)
16. Telgarsky, M., Vattani, A.: Hartigan's method: *k*-means clustering without Voronoi. In: Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 820–827 (2010)

17. Nielsen, F., Boissonnat, J.D., Nock, R.: On Bregman Voronoi diagrams. In: ACM-SIAM Symposium on Discrete Algorithms, pp. 746–755 (2007)
18. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Networks **16**(3), 645–678 (2005)
19. Kulis, B., Jordan, M.I.: Revisiting k-means: new algorithms via Bayesian nonparametrics. In: International Conference on Machine Learning (ICML) (2012)
20. Ackermann, M.R.: Algorithms for the Bregman K-median problem. PhD thesis. Paderborn University (2009)
21. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
22. Ji, S., Krishnapuram, B., Carin, L.: Variational Bayes for continuous hidden Markov models and its application to active learning. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 522–532 (2006)
23. Hidot, S., Saint-Jean, C.: An Expectation-Maximization algorithm for the Wishart mixture model: application to movement clustering. Pattern Recogn. Lett. **31**(14), 2318–2324 (2010)
24. Brent. R.P.: Algorithms for Minimization Without Derivatives. Courier Dover Publications, Mineola (1973)
25. Bezdek, J.C., Hathaway, R.J., Howard, R.E., Wilson, C.A., Windham, M.P.: Local convergence analysis of a grouped variable version of coordinate descent. J. Optim. Theory Appl. **54**(3), 471–477 (1987)
26. Bogdan, K., Bogdan, M.: On existence of maximum likelihood estimators in exponential families. Statistics **34**(2), 137–149 (2000)
27. Ciuperca, G., Ridolfi, A., Idier, J.: Penalized maximum likelihood estimator for normal mixtures. Scand. J. Stat. **30**(1), 45–59 (2003)
28. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
29. Nielsen, F.: Closed-form information-theoretic divergences for statistical mixtures. In: International Conference on Pattern Recognition (ICPR), pp. 1723–1726 (2012)
30. Haff, L.R., Kim, P.T., Koo, J.-Y., Richards, D.: Minimax estimation for mixtures of Wishart distributions. Ann. Stat. **39**(6), 3417–3440 (2011)
31. Jebara, T., Kondor, R., Howard, A.: Probability product kernels. J. Mach. Learn. Res. **5**, 819–844 (2004)
32. Moreno, P.J., Ho, P., Vasconcelos, N.: A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications. In: Advances in Neural Information Processing Systems (2003)
33. Petersen, K.B., Pedersen, M.S.: The matrix cookbook. http://www2.imm.dtu.dk/pubdb/p.php?3274. Accessed Nov 2012