Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Mean shift algorithm for exponential families with applications to speaker clustering

Themos Stafylakis[1,2], *Vassilis Katsouros*[3], *Patrick Kenny*[1,2], *Pierre Dumouchel*[1,2]

École de Technologie Supérieure (ÉTS), Quebec, Canada
Centre de Recherche Informatique de Montréal (CRIM), Quebec, Canada
Institute for Language and Speech Processing (ILSP), "Athena" R.C., Athens, Greece
{themos.stafylakis, patrick.kenny, pierre.dumouchel}@crim.ca
vsk@ilsp.athena-innovation.gr

Abstract

This work extends the mean shift algorithm from the observation space to the manifolds of parametric models that are formed by exponential families. We show how the Kullback-Leibler divergence and its dual define the corresponding affine connection and propose a method for incorporating the uncertainty in estimating the parameters. Experiments are carried out for the problem of speaker clustering, using both single Gaussians and i-vectors.

1. Introduction

Mean shift (MS) is a nonparametric clustering algorithm that has become a milestone in several areas of computer vision and image processing, [1]. Its main strengths are the mild assumptions that are required about the shapes of the clusters and the automatic estimation of their cardinality, [2]. It does so by finding the modes of the nonparametrically estimated density function and assigning each observation according to the basin of attraction of each mode.

What restricts us though from applying the MS to more general problems is the use of the (squared) Euclidean distance into the kernel. There are several tasks where either the observations or the parameters we use to encode them do not lie on $\Re^d$ and follow a non-Euclidean geometry. The subset of these problems that the paper deals with is the broad class of exponential families. Consider for example objects that are usually parametrized via histograms. In such cases, the objects lie on the manifold of multinomial distributions, which has at least two well-known geometries; those that are determined by the two Kullback-Leibler (KL) divergences, [3], [4]. Other approaches to make the MS algorithm applicable to manifolds may be found in [5] and [6]. Contrary to these approaches, we utilize the intrinsic structures of *statistical* manifolds and rely on the well-known KL divergences, instead of geodesics.

A first version of this paper was presented in [7]. The method is enhanced using MAP-estimation and a more precise mathematical derivation. The rest of the paper is organized as follows. In Sect. 2, the baseline MS algorithm is reviewed, along with the fundamental properties of exponential families. In Sect. 3, the proposed set-up is demonstrated, including the proposed kernels and the smoothing term. In Sect. 4, the proposed algorithm is derived, while a set of experiments is given in Sect. 5.

2. Preliminaries

2.1. The baseline mean shift algorithm

Let us first review the baseline mean shift algorithm. Consider a set of observations denoted by $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$, $\mathbf{x}^{(i)} \in \Re^d$ that have been generated by an unknown density $f(\mathbf{x})$. A nonparametric estimate $\hat{f}(\mathbf{x})$ of $f(\mathbf{x})$ is given by the following formula

$$\hat{f}_{h,k}(\mathbf{x}) = \frac{c_{k,d}}{Nh^d} \sum_{i=1}^{N} k\left(\left\| \frac{\mathbf{x} - \mathbf{x}^{(i)}}{h} \right\|^2 \right), \quad (1)$$

where $k(\cdot)$ the functional form of the unnormalized kernel, $c_{k,d}$ the inverse normalizing constant for unitary bandwidth, and h the bandwidth which controls the amount of smoothing. Note that h may vary with i, leading to the variable-bandwidth MS algorithm, [8].

The MS algorithm aims to assign the observations to clusters using a simple heuristic idea. It estimates the modes of the unknown density by setting the gradient of (1) with respect to $\mathbf{x}$ equal to zero. The gradient is as follows

$$\nabla \hat{f}_{h,k}(\mathbf{x}) = \frac{2c_{k,d}}{Nh^{d+2}} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \mathbf{x}) g\left(\left\| \frac{\mathbf{x} - \mathbf{x}^{(i)}}{h} \right\|^2 \right), \quad (2)$$

where $g(x) = -k'(x)$ and $x = \left\| \frac{\mathbf{x} - \mathbf{x}^{(i)}}{h} \right\|^2$. Note that $g(x) = \frac{1}{2}k(x)$ if the normal kernel

$$k_{\mathcal{N}}(x) = \exp\left(-\frac{1}{2}x \right) \quad (3)$$

is deployed. By placing the differential kernel $g(x)$ in (2) and rearranging some terms, we end-up with the following expression

$$\hat{\nabla} f_{h,k}(\mathbf{x}) = \frac{2c_{g,d}}{h^2 c_{g,d}} \hat{f}_{h,g}(\mathbf{x}) \mathbf{m}_{h,g}(\mathbf{x}), \quad (4)$$

where the two terms are as follows,

$$\hat{f}_{h,g}(\mathbf{x}) = \frac{2c_{g,d}}{Nh^{d+2}} \sum_{i=1}^{N} g\left(\left\| \frac{\mathbf{x}^{(i)} - \mathbf{x}}{h} \right\|^2 \right) \quad (5)$$

and

$$\mathbf{m}_{h,g}(\mathbf{x}) = \frac{\sum_{i=1}^N \mathbf{x}^{(i)} g\left(\left\| \frac{\mathbf{x}^{(i)} - \mathbf{x}}{h} \right\|^2 \right)}{\sum_{i=1}^N g\left(\left\| \frac{\mathbf{x}^{(i)} - \mathbf{x}}{h} \right\|^2 \right)} - \mathbf{x}. \quad (6)$$

The term $\mathbf{m}_{h,g}(\mathbf{x})$ is the mean shift vector i.e. the main result of the analysis. It points to the direction of maximum increase of $\hat{f}_{h,K}(\mathbf{x})$, given its current position $\mathbf{x}$. As (6) shows, the next position is a simple weighted average of the observations $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$, with the ith weight being equal to the proximity between $\mathbf{x}^{(i)}$ the current position $\mathbf{x}$, measured with the kernel profile $g(x)$.
The MS algorithm is as follows. For each observation $i = 1, 2 \ldots, N$ set $j = 0$, $\mathbf{x}_j \leftarrow \mathbf{x}^{(i)}$

1. calculate $\mathbf{m}_{h,G}(\mathbf{x}_j)$
2. set $\mathbf{x}_{j+1} \leftarrow \mathbf{x}_j + \mathbf{m}_{h,G}(\mathbf{x}_j)$
3. if $\|\mathbf{x}_{j+1} - \mathbf{x}_j\| < \epsilon$ goto 4; else $j \leftarrow j + 1$ and goto 1.
4. store $\tilde{\mathbf{x}}^{(i)} = \mathbf{x}_{j+1}$.

The matrix $\mathbf{X}_c = [\mathbf{x}_c^{(1)}, \mathbf{x}_c^{(2)}, \ldots, \mathbf{x}_c^{(N)}]$ contains the points that each observation converged. We only need to group those points having identical values, or more realistically those that the one-by-one distances do not exceed a small threshold ϵ.

2.2. Exponential families and their fundamental properties

Exponential families are a broad class of distributions $\mathcal{F}_\psi$ with certain appealing properties that allow us to treat them in a unified framework. Let us assume a d-dimensional observation vector $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T$ lying in some space $\mathcal{X} \subseteq \Re^d$. Let also $\mathbf{t}(\mathbf{x}) = \{t_\alpha(\mathbf{x})\}_{\alpha=1}^{n}$ be a set of functions $t_\alpha : \mathcal{X} \mapsto \Re$, known as *sufficient statistics*. Finally, let $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \Re^n$ the vector that contains the *natural* of *canonical* parameters, by which the distribution is parametrized. The probability density function (p.d.f.) p_ψ of the distribution $\mathcal{F}_\psi$ is expressed as follows

$$p_\psi(\mathbf{x}; \boldsymbol{\theta}) = \exp(\langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{x}) \rangle - \psi(\boldsymbol{\theta})), \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes dot-product. We emphasize that the p.d.f. in (7) is expressed with respect to (w.r.t.) an appropriate base measure $d\nu$ (not necessarily the Lebesgue $d\mathbf{x}$), such that for any measurable set $\mathbf{S}$, $P[\mathbf{x} \in \mathbf{S}] = \int_{\mathbf{S}} p_\psi(\mathbf{x}; \boldsymbol{\theta}) \nu(d\mathbf{x})$, [9].
An exponential family is *regular* if $\boldsymbol{\Theta}$ is an open set, and its representation is *minimal* if there is no nonzero vector $\mathbf{a} \in \Re^n$ which makes $\langle \mathbf{a}, \mathbf{t}(\mathbf{x}) \rangle$ a constant.
The function $\psi(\cdot) : \boldsymbol{\Theta} \mapsto \Re$

$$\psi(\boldsymbol{\theta}) = \log \int_{\mathcal{X}} \exp(\langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{x}) \rangle) \nu(d\mathbf{x}) \tag{8}$$

is called the *log partition* function and ensures that p_ψ is normalizable, i.e. $\int_{\mathcal{X}} \exp(\langle \boldsymbol{\theta}, \mathbf{t}(\mathbf{x}) \rangle - \psi(\boldsymbol{\theta})) \nu(d\mathbf{x}) = 1$. The log partition is strictly convex in $\boldsymbol{\Theta}$ and therefore $\boldsymbol{\Theta}$ is a convex convex set. Furthermore, $\boldsymbol{\Theta}$ and $\psi(\cdot)$ admit a dual space and function, respectively. First, we define the dual space of $\boldsymbol{\Theta}$ via by the gradient of $\psi(\boldsymbol{\theta})$, i.e. $\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})|_{\boldsymbol{\theta}}$. The dual parameters $\boldsymbol{\eta} \in \mathbf{H} \subseteq \Re^n$ are termed *expectation* parameters, since $\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{\theta}}[\mathbf{t}(\mathbf{x})]$, where $\mathbf{E}_{\boldsymbol{\theta}}[g(\mathbf{x})]$ is the expectation operator. The dual function is defined as

$$\phi(\boldsymbol{\eta}(\boldsymbol{\theta})) = \mathbf{E}_{\boldsymbol{\theta}}[\log p_\psi(\mathbf{x}; \boldsymbol{\theta})], \tag{9}$$

i.e. is simply the negative entropy of $p_\psi(\cdot; \boldsymbol{\theta})$. The inverse mapping $\mathbf{H} \mapsto \boldsymbol{\Theta}$ is defined via the gradient of $\phi(\cdot)$, i.e. $\boldsymbol{\theta}(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}} \phi(\boldsymbol{\eta})|_{\boldsymbol{\eta}}$.
Let us denote by $d_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = d_\phi(\boldsymbol{\eta} \| \boldsymbol{\eta}')$ the Kullback-Leibler divergence, defined as $d_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = \mathbf{E}_{\boldsymbol{\theta}} \left[\log \frac{p_\psi(\mathbf{x}; \boldsymbol{\theta})}{p_\psi(\mathbf{x}; \boldsymbol{\theta}')} \right]$. For exponential families, it can be expressed as follows

$$d_\psi(\boldsymbol{\theta} \| \boldsymbol{\theta}') = \psi(\boldsymbol{\theta}') - \psi(\boldsymbol{\theta}) - \left\langle \boldsymbol{\theta}' - \boldsymbol{\theta}, \boldsymbol{\eta} \right\rangle. \tag{10}$$

Since $\psi(\cdot)$ is strictly convex in $\boldsymbol{\Theta}$, the $n \times n$ matrix $G(\boldsymbol{\theta}) = \nabla^2_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})|_{\boldsymbol{\theta}}$ is positive definite and equals to Fisher Information Matrix (FIM), defined as follows

$$G_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = -\mathbf{E}_{\boldsymbol{\theta}} \left[\nabla^2_{\boldsymbol{\theta}} \log p_\psi(\mathbf{x}; \boldsymbol{\theta}) \right]. \tag{11}$$

The role of $G_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and its dual $G_{\boldsymbol{\eta}}(\boldsymbol{\eta}(\boldsymbol{\theta})) = G_{\boldsymbol{\theta}}(\boldsymbol{\theta})^{-1}$ is fundamental in statistics (e.g. Cramer-Rao bound, Jeffreys prior) and information geometry. In the latter field, the FIM can be used as the metric tensor of the Riemannian manifold of probability distributions, and enables us to define lengths, angles, volume elements and covariant derivatives.

3. Estimating the density function

In this section, we propose a framework to estimate underlying p.d.f. over which the mean shift algorithm will operate. Like the original algorithm, the p.d.f. is expressed nonparametrically, by a weighted summation of N kernel (or kernel-like) functions, one for each object. The squared distances will be replaced by KL divergences, while the kernels will be replaced by prior density functions that are used in the Bayesian statistical framework.

3.1. The kernel-like functions

Let us assume that we are given a set of N object $\{p^{(i)}\}_{i=1}^{N}$, where $p = p_\psi(\cdot, \boldsymbol{\theta})$, all belonging to the same exponential family defined by $\psi(\cdot)$. We should emphasize that the choice of parametrization is arbitrary. Let $\boldsymbol{\varphi} \in \Phi \subseteq \Re^n$. For the multivariate Normal distribution, we may consider $\boldsymbol{\varphi} = (\boldsymbol{\mu}, \Sigma)$ or we may express these objects e.g. with the expectation parameters $\{p_{\boldsymbol{\eta}}^{(i)}\}_{i=1}^{N}$. The parametrization over which the MS algorithm will operate is completely determined by the choice of the divergence.
Like the original MS algorithm, we start by expressing the empirical density

$$f_{emp}(\boldsymbol{\varphi}) = \frac{1}{N} \sum_{i=1}^{N} \delta(\boldsymbol{\varphi}, \boldsymbol{\varphi}^{(i)}) \tag{12}$$

on an arbitrary parametrization, say $\boldsymbol{\varphi} \in \Phi$. The empirical density should be smoothed by using a kernel, or a kernel-like function, since KL divergences are not symmetric in general. Let us introduce the following family of kernel-like functions

$$\Pi_\alpha(\boldsymbol{\varphi}; \boldsymbol{\varphi}_0, \lambda) = \exp(-\lambda D_\alpha(p_{\boldsymbol{\varphi}} \| p_0)). \tag{13}$$

This family is parametrized by $\alpha = \{-1, +1\}$ which defines the divergence function as follows

$$\alpha = \begin{cases} -1: & \text{Kullback-Leibler} \\ +1: & \text{swapped Kullback-Leibler} \end{cases} \tag{14}$$

to be compatible with the α-divergence discussed in [3]. We should emphasize that the kernel-functions in (13) are unnormalized densities, expressed w.r.t. the *natural volume element* $dV = \sqrt{g(\boldsymbol{\varphi})} d\boldsymbol{\varphi}$, where $g(\boldsymbol{\varphi}) = |G_{\boldsymbol{\varphi}}(\boldsymbol{\varphi})|$, [10]. Finally, let $\xi_\alpha(\boldsymbol{\varphi}_0, \lambda) = \int_V \Pi_\alpha(\boldsymbol{\varphi}; \boldsymbol{\varphi}_0, \lambda) dV$ be the normalizer of (13).
In Bayesian terms, $(\lambda, \boldsymbol{\varphi}_0)$ correspond to the hyperparameters of the distribution, with λ being the (not necessarily integer) number of virtual observations and $\boldsymbol{\varphi}_0$ the centering parameter. The two extreme cases where $\lambda = 0$ and $\lambda \to +\infty$ correspond to the Jeffreys (i.e noninformative) prior on $\boldsymbol{\varphi}$ and to a point mass concentrated at $\boldsymbol{\varphi}_0$, respectively. The case where $\alpha = +1$ corresponds to the familiar conjugate (to the likelihood function $p_\psi(\cdot; \boldsymbol{\varphi})$) prior, while $\alpha = -1$ to the entropic prior, studied in [10].

3.2. Taking into account the uncertainty in the estimates

Treating $\{p^{(i)}\}_{i=1}^N$ as point masses is in line with the original algorithm, where we begin by smoothing the empirical distribution with a kernel. However, there are cases where $\{p^{(i)}\}_{i=1}^N$ are not directly observable, but are rather being estimated based on a finite amount of observations. In such cases, we work as follows. Let us first consider the conjugate case $\alpha = +1$ and let $\hat{\boldsymbol{\varphi}}^{(i)} = \boldsymbol{\varphi}(\bar{\mathbf{t}}(\mathbf{X}_i))$ denote the ML estimate of $\boldsymbol{\varphi}^{(i)}$ given a sample $\mathbf{X}_i = \{\mathbf{x}^{(i,j)}\}_{j=1}^{n_i}$ expressed in an arbitrary $\boldsymbol{\varphi}$-parametrization. Due to the conjugacy to the likelihood, the density of the form of $\Pi_{+1}(\boldsymbol{\varphi}; \hat{\boldsymbol{\varphi}}^{(i)}, \lambda^{(i)})$, with $\lambda^{(i)} = n_i$ is the posterior of $\boldsymbol{\theta}$, using a flat (i.e. Jeffreys) prior. To incorporate our prior knowledge in the estimation, we attach an informative prior to $\boldsymbol{\varphi}^{(i)}$, using a set of n_0 virtual observations having expectation parameters equal to $\boldsymbol{\eta}_0$. In this case, the ML estimate will be replaced by $\tilde{\boldsymbol{\varphi}}^{(i)}$, defined as follows

$$\tilde{\boldsymbol{\varphi}}^{(i)} = \boldsymbol{\varphi}\left(\frac{1}{n_i + n_0}(n_i \bar{\mathbf{t}}(\mathbf{X}_i) + n_0 \boldsymbol{\eta}_0)\right), \tag{15}$$

i.e. the weighted barycenter of $\bar{\mathbf{t}}(\mathbf{X}_i)$ and $\boldsymbol{\eta}_0$, expressed in the $\boldsymbol{\varphi}$-coordinates, or simply $\tilde{p}^{(i)}$ in the coordinate-free notation. Hence, instead of the empirical distribution (i.e. a summation of delta functions) we begin with a mixture of posterior distributions, each centered at $\tilde{p}^{(i)}$. Note that a more general notation would include different number of virtual observations per parameter, as in the case of Gaussian - inverse Gamma conjugate priors attached to (μ, σ^2).

However, the mixture of posteriors, despite of not being a mixture of delta functions (due to the finite sample sizes), still requires further smoothing in order to be regarded as an approximation to the underlying p.d.f. of $\{p^{(i)}\}_{i=1}^N$. This is because we assume that $\{n_i\}_{i=1}^N \to +\infty$ is equivalent to the original algorithm, assuming objects in $\Re^d$ and the use of euclidean geometry. Note, therefore, that the proposed method does not assume that the estimates of two or more objects that belong to the same class converge asymptotically to the same point. On the contrary, a nonparametric discrepancy is allowed, in order to capture several other types of intra-class variability, that is discussed in more details in Sect. 4.2.

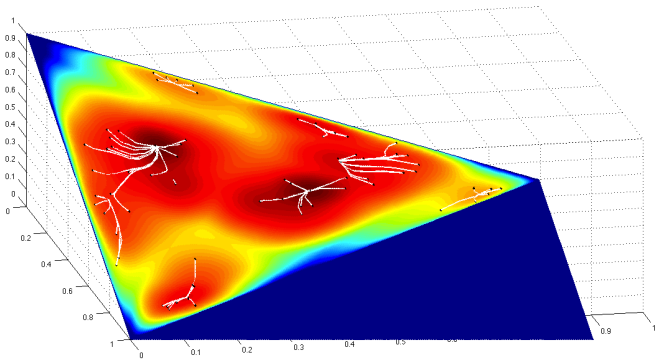

Figure 1: An example from the 3-dimensional multinomial family. The objects lie on the 2-simplex, the density is estimated using the $\alpha = +1$ kernel, namely the Dirichlet density, while the overall p.d.f. is depicted in the logarithmic scale using the heat colormap. The black markers show the set of $\{\tilde{p}^{(i)}\}_{i=1}^N$ while the white curves are their trajectories until they converge to their mode.

To apply this further smoothing, let $\tilde{n}_i = n_i + n_0$ and λ_0 be the upper bound of the precision, i.e. the inverse squared bandwidth. By adding the inverse precision λ_0^{-1} to the inverse precision of the posterior $\tilde{n}_i^{-1}$, we end-up with $\tilde{\lambda}^{(i)} = \frac{\lambda_0 \tilde{n}_i}{\lambda_0 + \tilde{n}_i}$. Therefore, the overall smoothed estimate of the underlying density will take the following expression

$$\tilde{f}_\alpha(p) = \frac{1}{N}\sum_{i=1}^N \xi_\alpha(\tilde{p}^{(i)}, \tilde{\lambda}^{(i)})^{-1} \Pi_\alpha\left(p; \tilde{p}^{(i)}, \tilde{\lambda}^{(i)}\right). \tag{16}$$

Note that the ML estimate $\hat{f}_\alpha(p)$ is fully recovered by $\tilde{f}_\alpha(p)$ when no uncertainty is assumed in estimating $\{p^{(i)}\}_{i=1}^N$, i.e. when $n_i \to \infty$, $i = 1, \ldots, N$.

4. The proposed mean shift algorithm

Having covered much of the theoretical background and the proposed method for estimating $\tilde{f}_\alpha(p)$, we demonstrate here how to adapt the MS algorithm to be compatible to our problem.

4.1. Deriving the mean shift iteration

The necessary condition for p to be a mode is to satisfy $\nabla \tilde{f}_\alpha(p) = \mathbf{0}$, i.e. to have zero gradient. Due to linearity we obtain

$$\frac{1}{N}\sum_{i=1}^N \nabla_{\boldsymbol{\theta}}\left[\xi_\alpha(\tilde{p}^{(i)}, \tilde{\lambda}^{(i)})^{-1} \Pi_\alpha\left(p; \tilde{p}^{(i)}, \tilde{\lambda}^{(i)}\right)\right] = \mathbf{0}. \tag{17}$$

Therefore, we obtain

$$\frac{1}{N}\sum_{i=1}^N \xi_\alpha(\tilde{p}^{(i)}, \tilde{\lambda}^{(i)})^{-1} \Pi_\alpha\left(p; \tilde{p}^{(i)}, \tilde{\lambda}^{(i)}\right) \times \left[-\tilde{\lambda}^{(i)} \nabla_{\boldsymbol{\theta}} D_\alpha(p \| \tilde{p}^{(i)})\right] = \mathbf{0} \tag{18}$$

Let us consider the $\alpha = \pm 1$ cases. For $\alpha = +1$, we obtain

$$\nabla_{\boldsymbol{\theta}} D_{+1}(p \| \tilde{p}^{(i)}) = \boldsymbol{\eta} - \tilde{\boldsymbol{\eta}}^{(i)} \tag{19}$$

where $\boldsymbol{\eta}$ and $\tilde{\boldsymbol{\eta}}^{(i)}$ are shorthands to $\boldsymbol{\eta}(p)$ and $\boldsymbol{\eta}\left(\tilde{p}^{(i)}\right)$, respectively. Therefore, the MS iteration for the $\alpha = +1$ case is as follows

$$\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} + \mathbf{m}_{\lambda_0, +1}, \tag{20}$$

where

$$\mathbf{m}_{\lambda_0,+1}(p) = \frac{\sum_{i=1}^N \tilde{\boldsymbol{\eta}}^{(i)} w_{\lambda_0,+1}^{(i)}(p)}{\sum_{i=1}^N w_{\lambda_0,+1}^{(i)}(p)} - \boldsymbol{\eta} \tag{21}$$

and

$$w_{\lambda_0,+1}^{(i)}(p) = \tilde{\lambda}^{(i)} \xi_{+1}(\tilde{p}^{(i)}, \tilde{\lambda}^{(i)})^{-1} \Pi_{+1}\left(p; \tilde{p}^{(i)}, \tilde{\lambda}^{(i)}\right). \tag{22}$$

For the $\alpha = -1$ case we obtain

$$\nabla_{\boldsymbol{\theta}} D_{-1}(p \| \tilde{p}^{(i)}) = G_{\boldsymbol{\theta}}(\boldsymbol{\theta})\left[\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}^{(i)}\right]. \tag{23}$$

Therefore, the MS iteration is as follows

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \mathbf{m}_{\lambda_0, -1}, \tag{24}$$

where $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}^{(i)}$ are shorthands for $\boldsymbol{\theta}(p)$ and $\boldsymbol{\theta}\left(\tilde{p}^{(i)}\right)$, respectively. Moreover,

$$\mathbf{m}_{\lambda_0,-1}(p) = \frac{\sum_{i=1}^N \tilde{\boldsymbol{\theta}}^{(i)} w_{\lambda_0,-1}^{(i)}(p)}{\sum_{i=1}^N w_{\lambda_0,-1}^{(i)}(p)} - \boldsymbol{\theta} \tag{25}$$

and

$$w^{(i)}_{\lambda_0,-1}(p) = \tilde{\lambda}^{(i)}\xi_{-1}(\tilde{p}^{(i)},\tilde{\lambda}^{(i)})^{-1}\Pi_{-1}\left(p;\tilde{p}^{(i)},\tilde{\lambda}^{(i)}\right). \quad (26)$$

Note that in order to derive (25) we used the fact that $G_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is positive-definite and therefore we may multiply both sides of an equation by its inverse. Note that if $\tilde{\lambda}^{(i)}$ is below a critical value, $\xi_\alpha(\tilde{p}^{(i)},\tilde{\lambda}^{(i)}) \to +\infty$. In such cases the normalizer can be omitted from the equations, and proceed with unnormalized kernels, as in [5]. An example of the proposed MS algorithm is illustrated in Fig. 1, for the 3-dimensional multinomial family, while an example on the bivariate normal is depicted in 2.

4.2. Discusion

A key issue that may cause misconceptions is the use of a parametric interpretation of the objects $\{p^{(i)}\}_{i=1}^N$ within a nonparametric algorithm. We should reemphasize though that the term nonparametric can be misleading, since it only refers to models that allow the number of parameters grow with the number of observations in a linear (e.g. kernel density estimates) or sublinear rate (e.g. Dirichlet process mixture models). On the contrary, what we propose is the use of a parametric descriptor for each object - which we further demand to be an exponential family. As an example, consider the following Normal-Inverse Wishart hierarchical model

$$(\mu_k,\Sigma_k) \sim \mathcal{NIW}(\mu_0,\Sigma_0,\alpha_0) \quad (27)$$

$$(\mu^{(i)},\Sigma^{(i)}) \sim P_{np}(\mu_k,\Sigma_k) \quad (28)$$

$$\{\mathbf{x}_i^{(j)}\}_{j=1}^{n_i} \sim \mathcal{N}(\mu^{(i)},\Sigma^{(i)}) \quad (29)$$

where $P_{np}(\mu_k,\Sigma_k)$ a non-parametric distribution with conditions of the first two moments. This model does assume that the data of the ith object $\{\mathbf{x}_i^{(j)}\}_{j=1}^{n_i}$ is sampled from a parametric model $\phi^{(i)} = (\mu^{(i)},\Sigma^{(i)})$. However, by adding an intermediate layer that introduces a nonparametric discrepancy, it allows the class-conditional distribution to vary across objects of the ith class around (μ_k,Σ_k) instead of being fixed. This discrepancy allows us to consider cases where classes may exhibit smooth, yet arbitrary shapes on the space of parametric distributions, caused by one or more types of variability (e.g. channel variability, [11]) and underlines the rationale for the proposed algorithm and its wide applicability. Several other methods used either ignore this hierarchical underline structure and apply heuristics rules to compensate for their simplifications (like the artificial boosting of the penalty term of complexity criteria like BIC) or deploy a parametric discrepancy distribution (e.g. a conjugate prior), even in cases where the data does not support such an assumption, [12].

A further misconception may be caused by the use of the proposed kernels. We should emphasize that in the same way the use of a Gaussian kernel in the original MS does not imply any assumption about the gaussianity of the class-conditional densities of the observations, the use of e.g. a conjugate prior as a kernel does not imply any assumption about the class-conditional densities of the objects. They are deployed in order to smooth the empirical distribution and not to model class-conditional densities.

In cases where the objects are too complex to be modeled accurately by an exponential family, one may consider mixtures models, e.g. a Gaussian mixture model. To make mixture models applicable for the proposed algorithm, one should estimate their parameters and latent variables $Z_i = \{z^{(i,j)}\}_{j=1}^{n_i}$ using a MAP-adaptation scheme. A prior distribution should first be estimated using enrollment data, including both the prior of each of the components and the prior for the weights (typically Dirichlet), and apply MAP-EM to estimate $\{\tilde{p}_i\}_{i=1}^N$ and the latent variables. Then, one may notice that the complete-data likelihood of mixture models belongs to an exponential family, iff the likelihood of the components belongs to an exponential family as well, [3]. Note that in the mixture case, the use of MAP-estimates is required not only for its robustness, but also in order to establish a fixed association between mixture components of two or more objects. Finally, we emphasize that the use of the complete-data likelihood (by treating the estimates of the latent variables as observable) is one of the building blocks of recent state-of-the-art speaker recognition, [11].

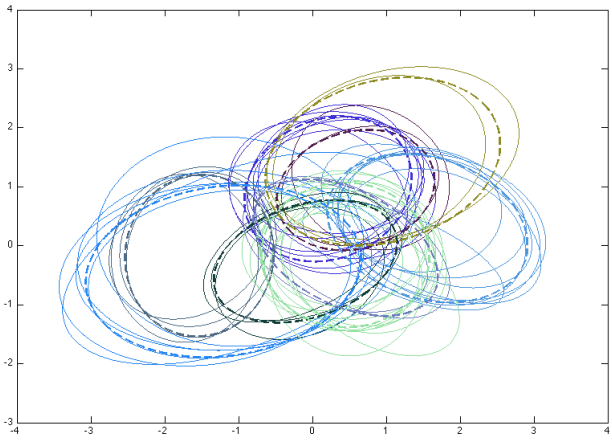

Figure 2: A two-dimensional example that depicts how the proposed algorithm successfully discovers the true partitioning of $N = 45$ Gaussian objects into $K = 9$ clusters, using the $\alpha = +1$ configuration. The objects are plotted via their corresponding ellipse, the color corresponds to the cluster label, while the estimated mode of each cluster is illustrated using bold-dashed line.

5. Experiments

We examine the strength of our algorithm with respect to the problem of speaker clustering, using the ESTER benchmark, [12]. The ESTER consists of 32 Broadcast News shows extracted from several French Radio channels, and is divided into the development (14 shows) and the test set (18 shows).

The features are 18-dimensional static MFCC (100Hz frame rate, augmented by the log-energy, while c_0 is discarded). An HMM with 128-component GMMs for each audio macro class is used in order to classify the frames into speech, silence, music and speech over music. A BIC-based speaker change detector is then applied on the stream. Finally, we evaluate the system using two different models. One where each speech segment is described via a single Gaussian distributions with a full covariance matrix and a second one using *i-vectors*, [11]. All user-defined parameters are tuned based on the development set.

5.1. Diarization using single Gaussian distributions

We compare our algorithm against the baseline ΔBIC-based Agglomerative Hierarchical Clustering (AHC), described in

[13], up to the Viterbi resegmentation stage. To score the methods, the official Diarization Error Rate (DER, %) metric is used. The metric is defined as the Hamming distance between the estimated and the reference clustering, plus the False Alarm (FA) and Missed Detection (MD) error rates.
Note that the expectation and the natural parameters for the Multivariate Gaussian distribution are as follows

$$\boldsymbol{\eta} = \left(\mu,\, \Sigma + \mu\mu^T\right),\ \boldsymbol{\theta} = \left(\Sigma^{-1}\mu,\, -\frac{1}{2}\Sigma^{-1}\right), \qquad (30)$$

where (μ, Σ) denote mean and covariance matrix, respectively.

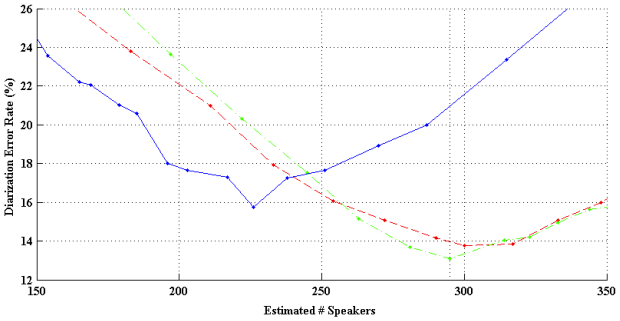

Figure 3: Estimated number of speakers vs. DER(%) for varying λ_0, on the ESTER development set. Blue curve: AHC with ΔBIC. Green curve: MS algorithm with $\alpha = +1$. Red curve: MS algorithm with $\alpha = -1$. The vertical line indicates the overall number of speakers in the set. The curves are drawn for various values of λ_0. For the AHC algorithm, λ_0 corresponds to the artificial boosting of the penalty term.

As Fig. 3 illustrates, the proposed algorithm has better performance compared to the baseline AHC, especially when considering the joint DER - ENS metric score. Moreover, the MS with $\alpha = +1$ showed increased performance compared to $\alpha = -1$. The best results for both sets can be found in Table 1. Note that by applying Viterbi resegmentation to the best MS configuration, the DER attained 13.29% on the test set, compared to 15.37% using the ΔBIC-based AHC. The results of the official ESTER evaluation are given in [12] and show that the proposed method is second, only behind the LIMSI GMM-UBM approach that scored 11.5%. Finally, we should note that the original MS algorithm, that is MS with only the mean values as parameters, scored an 18.62% DER. This rather poor score clearly demonstrates the need to extend it to more general non-Euclidean manifolds, that encode e.g. second order statistical information as well.
After tuning on the development set, we use $n_0 = 130$ (that corresponds to $1.3s$ duration), and center the prior of the covariance matrix equal to 0.75 of the averaged covariance per file, estimated on the development set. We observed that ML estimates work better when estimating the mean values, and therefore the results are derived using a flat prior over the means. The optimal λ_0 was found to be equal to 1.2 and 1.3 for $\alpha = +1$ and $\alpha = -1$, respectively. Since both λ_0 are below the critical value that makes $\Pi_\alpha(\cdot;\cdot,\cdot)$ normalizable, the normalizers $\xi_\alpha(\cdot,\cdot)$ were excluded from our equations. Note that even in cases where $\{n_i\}_{i=1}^N \gg \lambda_0$, the method still captures the information about sample sizes through the use of MAP-estimates. Finally, the objects (segments of speech in our case) required between 3 and 7 MS iteration to attain convergence.

Table 1: Minimum Overall Speaker Diarization Error Rate (%) for the two ESTER sets, using Gaussians

	ESTER-DEV	ESTER-TEST
AHC, ΔBIC	15.76	16.28
MS, $\alpha = +1$	**13.12**	**14.51**
MS, $\alpha = -1$	13.79	14.66
FA Rate	0.3	0.6
MD Rate	0.9	1.2

5.2. Diarization using i-vectors

We trained a gender-independent GMM UBM containing 512 Gaussians, using ESTER phase-II for enrollment data (about 100 hours overall duration). We use a gender independent i-vector extractor of dimension 400, trained on the same set. Cepstral mean subtraction is applied, using sliding windows. The i-vectrors are preprocessed using Linear Discriminant Analysis (LDA), and the dimensionality reduces to 200. Finally, Within-Class Covariance Normalization (WCCN) is applied.
What is interesting with i-vectors is the fact that they lie on the unit-sphere. Therefore, the corresponding exponential family is the Von Mises-Fisher distribution. The distribution retains the structure of the Euclidean distance, and uses a measure that places non-zero probability mass only to the surface of the unit sphere, [3]. Therefore, the two connection coincide with each other, like in the Euclidean case.
The algorithm is very similar to the original MS algorithm. The only difference is the fact that after each iteration, the new position is forced to lie on the surface by dividing by the norm. Moreover, since i-vectors are by themselves MAP estimates, there is no need to define any explicit center for the prior. However, the size of the segments is used in order the define the bandwidths, that converge to λ_0 as $\{n_i\}_{i=1}^N$ grow. In fact, we use an extra parameter $0 < r < 1$ to define $\tilde{\lambda}^{(i)}$ as follows

$$\tilde{\lambda}^{(i)} = \frac{n^{(i)} r \lambda_0}{n^{(i)} r + \lambda_0}, \qquad (31)$$

This can be explained by the fact that mfcc are not i.i.d., but exhibit strong autocorrelation. Thus, $n^{(i)}r$ may be considered as the effective sample size of the ith speech segment. Moreover, since $\lambda_0 \ll \{n_i\}_{i=1}^N$, r allows us to encode the uncertainty in the estimates.
The algorithm is compare to a hierarchical clustering one, that used a fixed threshold to decide whether two segments should be merged or not. After merging, the new cluster is defined as the weighted average of the i-vectors in $\Re^d$, and then projected onto the unit sphere. The weights are proportional to the sizes of the segments.
The results are demonstrated in Table 2, and show that the AHC outperformed MS by a small margin. We are planning to enhance the results by using standard non-parametric approaches, with the variable-bandwidth MS being the most prominent one.

6. Conclusion and future work

This paper proposed an extension of the MS algorithm for exponential families. We showed how the core idea of the algorithm can be applied to the particular manifolds and how the choice of the KL divergence determines the affine coordinates for each case. An extension to deal with objects that are not directly ob-

Table 2: Minimum Overall Speaker Diarization Error Rate (%) for the two ESTER sets, using i-vectors

	ESTER-DEV	ESTER-TEST
AHC	**10.12**	**12.18**
MS	11.42	13.71
FA Rate	0.3	0.6
MD Rate	0.9	1.2

servable is also given.
Our future work includes a variety of directions, such as the use of other kernel-like functions and divergences (e.g. Hellinger distance, $\alpha = 0$), the adaptation of methods that automatically estimate the tuning parameter λ_0, as well as further experiments on several other machine learning tasks.

7. Aknowledgments

This work has been partly financed by the Greek Secretariat for Research & Technology through project WELCOM-09SYN-71-856 (NSRF 2007-2013).

8. References

[1] D. Comaniciu and P. Meer, "Mean shift: A robust approach towards feature space analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603 – 619, May 2002.

[2] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. on Information Theory*, vol. 21, no. 1, pp. 32–40, January 1975.

[3] S. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, vol. 8, pp. 1379–1408, 1995.

[4] ——, "Differential Geometry of Curved Exponential Families-Curvatures and Information Loss," *The Annals of Statistics*, 1982.

[5] O. Tuzel, R. Subbarao, and P. Meer, "Simultaneous multiple 3D motion estimation via mode finding on Lie groups," in *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 18–25.

[6] A. Shamir, L. Shapira, and D. Cohen-Or, "Mesh analysis using geodesic mean-shift," *Vis. Comput.*, vol. 22, no. 2, pp. 99–108, Feb. 2006.

[7] T. Stafylakis, V. Katsouros, and G. Carayannis, "Speaker clustering via the mean shift algorithm," in *Odyssey 2010: The Speaker and Language Recognition Workshop - Odyssey-10, Brno, Czech Republic*, June 2010.

[8] D. Comaniciu, V. Ramesh, and P. Meer, "The variable bandwidth Mean Shift and data-driven scale selection," in *Proc. 8th Intl. Conf. on Computer Vision*, 2001, pp. 438–445.

[9] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*. Hanover, MA, USA: Now Publishers Inc., 2008.

[10] C. C. Rodriguez, "Entropic priors for discrete probabilistic networks and for mixtures of gaussians models," in *Bayesian Inference and Maximum Entropy Methods*. Inst. Physics, 2001, pp. 410–432.

[11] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification." *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[12] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," in *Proceedings of European Conference on Speech Communication and Technology (Interspeech)*, September 2005, pp. 1149 – 1152.

[13] P. Deleglise, Y. Esteve, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based System for French Broadcast News," in *Proceedings of Interspeech, Lisbon, Portugal*, 2005.