Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Speaker clustering via the mean shift algorithm

Themos Stafylakis[1,2], *Vassilis Katsouros*[1], *George Carayannis*[1,2]

[1]Institute for Language and Speech Processing, Athena Research Center, Greece
[2]National Technical University of Athens, Greece
{themosst,vsk,gcara}@ilsp.athena-innovation.gr

Abstract

In this paper, we investigate the use of the mean shift algorithm with respect to speaker clustering. The algorithm is an elegant nonparametric technique that has become very popular in image segmentation, video tracking and other image processing and computer vision tasks. Its primary aim is to detect the modes of the underlying density and consequently merge those observations being attracted by each mode. Since the number of modes is not needed to be known beforehand, the algorithm seems to fit well to the problem of speaker clustering. However, the algorithm needs to be adapted; the original algorithm acts on the space of observations, while speaker clustering algorithms act on the space of probabilistic parametric models. We attempt to adapt the algorithm, based on some basic concepts of information geometry, that are related to the exponential family of distributions.

1. Introduction

The problem of speaker clustering plays a fundamental role in speech technologies, since is related to a variety of tasks, such as speech and speaker recognition, rich-transcription of dialogues and others, [1]. Given a collection of speech segments and assuming that for each of them one and only one speaker is active, the goal is to merge those segments being uttered by the same speaker. No target-speaker model should be used while the number of speakers should be estimated from the data.

When dealing with a unique audio file, such as a broadcast or a meeting, many implementations divide the task into two distinct subproblems. The first problem is called speaker segmentation and aims to chop the file into speaker turns, while the second one is the problem of speaker clustering. The overall procedure is entitled speaker diarization and the branch of algorithms that follow this paradigm are usually called step-by-step or disjoint, as opposed to the integrated ones, [2]. The latter branch does not apply any explicit segmentation into speaker turns, uses an HMM framework to avoid fast transitions between speakers and relies either on the frequentist framework (Evolving-HMMs) or to more concrete Bayesian models (Dirichlet processes) and inferential procedures (MCMC [3], Variational Bayes [4], a.o.). Hence, when dealing with audio files instead of collections of separate audio segments, the proposed method requires a segmentation stage to operate, and therefore should be regarded as an alternative to the dominant approach of agglomerative hierarchical clustering (AHC), [5].

The mean shift (MS) algorithm has gained a lot of attention amongst image and video processing communities. The applications that make use of it range from color and motion segmentation to discontinuity - preserving smoothing and tracking, [6]. Its elegancy arises from the way it bypasses the well-known problem of nonparametric density estimation of multivariate data, namely the sparsity of the data due to the curse of dimensionality, [7]. It starts by observing that a robust estimation of the underlying density is out of the scope of many machine learning tasks; for many such tasks, what suffices is the extraction of certain characteristics of the density. Considering the clustering task, estimating the modes of the unknown density and deriving a rule to assign each observation to the appropriate mode is what ultimately needed, at least for obtaining a point-estimate of the assignments. By restricting ourselves to hard clustering, it turns out that the mean shift algorithm offers both. The observations are assigned to a cluster via the *basin of attraction* that each mode creates round it. Moreover, the clusters may exhibit smooth yet arbitrary shapes in the feature space, due to the nonparametric setting. Finally, the number of modes is not required beforehand.

At a first glance, the algorithm seems to meet the demands of being applicable to speaker clustering. However, a main difference between speaker clustering and image segmentation is that in the former task, each segment is described by a parametric model (say a Gaussian Mixture Model, GMM, or a single Gaussian) i.e. the entities lie on the space of parameters of probabilistic models instead of the space of observations. It is clear the Euclidean geometry is no long useful and one needs to reform the algorithm accordingly. Nevertheless, thanks to the works of Amari ([8]) and many other researchers ([9], [10], [11], [12]), the geometry of parametric models has been examined in depth, under the term *Information Geometry*. One needs to assume that the parameter space is a Riemannian manifold and that the metric is defined by the Fisher Information Metric Tensor. Furthermore, the squared distance has its natural analogue to the Kullback-Leibler Divergence. Moreover, if we restrict our analysis to models that belong to the exponential family, the mathematical analysis becomes much easier; we can make use of the duality between the natural parameter vector and the expectation parameter vector.

The rest of the paper is organized as follows. In Sect. 2, the baseline mean shift algorithm is explained, while the basic theory of the exponential family is presented in Sect. 3. In Sect. 4, some of the main properties of the Kullback-Leibler Divergence are discussed, along with the proposed kernel. The proposed method is analyzed in Sect. 5, while the experiments are presented in Sect. 6. Finally, some technicalities of the implementation are discussed in Sect. 7, where possible extension are being considered, too.

2. The mean shift algorithm on the space of observations

In this section, we present the derivation of the mean shift algorithm into its original formulation. The derivation is more or less based on the seminal paper of Comaniciu & Meer, [13]. For completeness we revise some or their main results.

2.1. Nonparametric density estimation basics

Suppose we are given a collection of observations $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, $\mathbf{x}^{(i)} \in \Re^d$. Let us assume that the unknown density $f(\mathbf{x})$ that generates the data can be estimated by the following density

$$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}^{(i)}), \tag{1}$$

where

$$K_{\mathbf{H}}(\mathbf{x}) = c_d|\mathbf{H}|^{-1/2}K(\mathbf{H}^{-1/2}\mathbf{x}) \tag{2}$$

is the kernel of the method and $\mathbf{H}$ is a positive definite matrix. This is a typical nonparametric setting, where we smooth the empirical density

$$f_{emp}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\mathbf{x}, \mathbf{x}^{(i)}) \tag{3}$$

by convolving it with a kernel $K_{\mathbf{H}}(\mathbf{x})$ centered at the origin ($\delta(\cdot,\cdot)$ is the Kronecker delta function). Moreover, let us simplify the analysis and assume that $\mathbf{H} = h^2\mathbf{I}_d$, where $\mathbf{I}_d$ is the d-dimensional identity matrix, and consider only radically symmetric kernels,

$$K(\mathbf{x}) = c_{k,d}k(||\mathbf{x}||^2). \tag{4}$$

The function $k(\cdot)$ is known as the *profile* of the kernel, that takes as argument the scalar squared distance. A common choice is the Gaussian kernel, having profile the following function

$$k_N(x) = \exp\left(-\frac{1}{2}x\right), \tag{5}$$

that corresponds to the d-variate Gaussian kernel

$$K_N(\mathbf{x}) = (2\pi)^{d/2}\exp\left(-\frac{1}{2}||\mathbf{x}||^2\right). \tag{6}$$

By using the profile notation with a fixed bandwidth h, the density estimation function in (1) yields

$$\hat{f}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x} - \mathbf{x}^{(i)}}{h}\right\|^2\right). \tag{7}$$

Recall that the term nonparametric can be misleading; it corresponds to an estimation setting where the number of parameters is let to grow linearly with n. The parameters are the centers of the kernels (i.e. the observations) and the bandwidth h. Moreover, the use of a unique h is not linked to the nonparametric setting; variable bandwidth alternatives may also be examined, and usually enhance the robustness, with the cost of being more computationally demanding, [14]. If the variable bandwidth approach is deployed, the bandwidth should decrease as n grows, to obtain an asymptotically unbiased estimator $\hat{f}(\cdot)$ of $f(\cdot)$, and minimize the sum of variance and squared bias, [13].

2.2. Mode seeking via the gradient

The mean shift algorithm estimates the modes of the unknown density by setting the gradient of (7) with respect to $\mathbf{x}$ equal to zero. The gradient is as follows

$$\nabla\hat{f}_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(\mathbf{x} - \mathbf{x}^{(i)})k'\left(\left\|\frac{\mathbf{x} - \mathbf{x}^{(i)}}{h}\right\|^2\right). \tag{8}$$

By setting $g(x) = -k'(x)$ we introduce a second kernel profile, which equals to the negative derivative of $k(x)$ with respect to x. The corresponding kernel is denoted by $G(\mathbf{x})$ and has the following form

$$G(\mathbf{x}) = c_{g,d}g(\|\mathbf{x}\|^2). \tag{9}$$

Note that if the Gaussian kernel is used, the two profiles coincide. By placing the differential kernel $g(x)$ in (8) and rearranging some terms, we end-up with the following expression

$$\hat{\nabla} f_{h,K}(\mathbf{x}) = \frac{2c_{k,d}}{h^2 c_{g,d}}\hat{f}_{h,G}(\mathbf{x})\mathbf{m}_{h,G}(\mathbf{x}), \tag{10}$$

where the two terms are as follows,

$$\hat{f}_{h,G}(\mathbf{x}) = \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}^{(i)} - \mathbf{x}}{h}\right\|^2\right) \tag{11}$$

and

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbf{x}^{(i)}g\left(\left\|\frac{\mathbf{x}^{(i)}-\mathbf{x}}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}^{(i)}-\mathbf{x}}{h}\right\|^2\right)} - \mathbf{x} \tag{12}$$

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 (12) 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 $g(x)$.

2.3. The mean shift algorithm

Having fixed much of the theoretical background, we now present the algorithm that implements the above idea.
For each observation $i = 1, 2\ldots, n$ set $t = 0$, $\mathbf{x}_t \leftarrow \mathbf{x}^{(i)}$

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

The matrix $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}^{(1)}, \tilde{\mathbf{x}}^{(2)}, \ldots, \tilde{\mathbf{x}}^{(n)}]$ contains the points in $\Re^d$ 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 (say ϵ).
Its worth referring to the self-normalizing property of the resulting iterative procedure. Using (10), the mean shift vector can be expressed as follows

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{1}{2}h^2 c\frac{\hat{\nabla} f_{h,K}(\mathbf{x})}{\hat{f}_{h,G}(\mathbf{x})} \tag{13}$$

The denominator of the above expression demonstrates that in areas of low (high) densities (as estimated with kernel profile $g(x)$) the step will be large (small). This property is very appealing; the algorithm normalizes the magnitude of each step according to the density estimate $\hat{f}_{h,G}(\mathbf{x}_t)$ and hence it eliminates the need of user-defined parameters in order to stabilize or accelerate the process.
Finally, considering that for the Gaussian kernel as well as for its truncated versions, the two profiles are identical, (13) shows that the mean shift vector is proportional to the derivative of the estimated log-density at $\mathbf{x}$ with respect to $\mathbf{x}$, [7].

3. Exponential family fundamentals

In this section, we review some of the basic elements needed in order to make the mean shift algorithm compatible to our goal. We explain certain properties of the exponential family, a member of which is the d-variate Gaussian distribution.

3.1. Main properties of the exponential family

The exponential family consists of a broad class of distributions with certain appealing properties. Its corresponding density function has the following form

$$p(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})\exp(\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{x}) - \psi(\boldsymbol{\theta})) \tag{14}$$

where

$$\psi(\boldsymbol{\theta}) = \log\int_{\mathcal{X}}\exp(\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{x}))h(\mathbf{x})d\mathbf{x} \tag{15}$$

is the *log-partition* function (i.e. the logarithm of the normalizing constant) and $h(\mathbf{x})d\mathbf{x}$, $h : \mathcal{X} \mapsto R^+$ the reference measure. Since $h(\mathbf{x})$ is constant for the case of Gaussians (when both mean and variance are unknown), it can be absorbed by $\psi(\boldsymbol{\theta})$ and we may simply set $h(\mathbf{x}) = 1$ in (14). Furthermore, $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^P$ denotes the P-dimensional vector of the *natural parameters*, and $\mathbf{t}(\mathbf{x})$ the vector of the *sufficient statistics* of $\mathbf{x}$, i.e. a map $\mathcal{X} \mapsto \Re^P$. For the univariate Gaussian distribution, the above functions have the following form

$$\boldsymbol{\theta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right), \ \mathbf{t}(x) = \left(x, x^2\right) \tag{16}$$

and

$$\psi(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \tag{17}$$

where (μ, σ^2) denote the mean and the variance, respectively. The log-partition function has a fundamental role; by differentiating $\psi(\boldsymbol{\theta})$ we obtain the *expectation* parameters $\boldsymbol{\eta}(\boldsymbol{\theta})$, i.e.

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \left(\mu, \sigma^2 + \mu^2\right) \tag{18}$$

for the univariate case. Moreover, the second order derivative yields the Fisher Information with respect to the natural parameters

$$G(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\eta} \tag{19}$$

which equals to

$$G(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2 + 2\sigma^4 \end{pmatrix} \tag{20}$$

for the univariate case. It gives the lower bound of the covariance matrix of any unbiased estimator $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ based on a unitary sample size. Furthermore, (19) shows that $G(\boldsymbol{\theta})$ equals to the Jacobian of the transform, since $\{G(\boldsymbol{\theta})\}_{ij} = \frac{\partial\eta_i}{\partial\theta_j}$. In terms of Riemannian Geometry, $G(\boldsymbol{\theta})$ is the metric tensor on a parameter manifold, [8]

We may consider $\boldsymbol{\eta}$ as the dual coordinate system of $\boldsymbol{\theta}$. This property arises from the convexity of $\psi(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ that establishes a one-to-one mapping between them and ensures that $G(\boldsymbol{\theta})$ is positive definite. This duality becomes more apparent by introducing the dual potential function

$$\phi(\boldsymbol{\eta}) = -\frac{1}{2}\log(2\pi e\sigma^2) \tag{21}$$

that generates the natural parameters as follows

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}}\phi(\boldsymbol{\eta}) \tag{22}$$

Note that the dual potential is equal to the (negative) Shannon entropy of the distribution. Likewise (19), we further obtain

$$G(\boldsymbol{\eta}) = G(\boldsymbol{\theta})^{-1} = \nabla_{\boldsymbol{\eta}}\nabla_{\boldsymbol{\eta}}\phi(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}}\boldsymbol{\theta} \tag{23}$$

the Fisher Information Matrix with respect to the expectation parameters. The one-to-one correspondence between the two coordinate systems is a consequence of the convexity of $\psi(\boldsymbol{\theta})$ in $\boldsymbol{\theta}$ and is called the *Legendre Transform*, defined by

$$\phi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}}\{\boldsymbol{\theta}\cdot\boldsymbol{\eta} - \psi(\boldsymbol{\theta})\} \tag{24}$$

and its dual expression

$$\psi(\boldsymbol{\theta}) = \max_{\boldsymbol{\eta}}\{\boldsymbol{\theta}\cdot\boldsymbol{\eta} - \phi(\boldsymbol{\eta})\} \tag{25}$$

where the potentials satisfy the identity

$$\psi(\boldsymbol{\theta}) + \phi(\boldsymbol{\eta}) = \boldsymbol{\theta}\cdot\boldsymbol{\eta}. \tag{26}$$

Based on the above statistical entities, one can derive the corresponding ones for the multivariate case. The natural and expectation parameters have as follows

$$\boldsymbol{\theta} = \left(\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}\right), \boldsymbol{\eta} = \left(\mu, \Sigma + \mu\mu^T\right) \tag{27}$$

while the potentials are equal to

$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log((2\pi)^d|\Sigma|) \tag{28}$$

and

$$\phi(\boldsymbol{\eta}) = -\frac{1}{2}\log((2\pi e)^d|\Sigma|). \tag{29}$$

Note that their second entries are square matrices, i.e. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\Theta}_2)$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{H}_2)$, where $\boldsymbol{\Theta}_2, \boldsymbol{H}_2 \in \Re^{P\times P}$ and P equals to the dimension of (μ, Σ), i.e. $P = d + d(d+1)/2$. The dot-product between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is carried out as follows

$$\boldsymbol{\theta}\cdot\boldsymbol{\eta} = Tr\{\boldsymbol{\theta}_1\boldsymbol{\eta}_1^T + \boldsymbol{\Theta}_2\boldsymbol{H}_2^T\} = \boldsymbol{\theta}_1^T\boldsymbol{\eta}_1 + Tr\{\boldsymbol{\Theta}_2\boldsymbol{H}_2^T\} \tag{30}$$

where $|A|$, A^T and $Tr\{A\}$ denote determinant, transpose and trace of A, respectively.

4. The Kullback-Leibler Divergence and a probabilistic kernel

4.1. The Kullback-Leibler Divergence

Having covered much of the required theoretical background, we now define the Kullback-Leibler Divergence (KLD). Let $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}^l$ be the natural parameters of two distributions with densities $Q^k = p(\mathbf{x};\boldsymbol{\theta}^k)$ and $Q^l = p(\mathbf{x};\boldsymbol{\theta}^l)$ of the same class and dimensionality. We denote by $D(Q^k||Q^l)$ the KLD between $\boldsymbol{\theta}^k$ and $\boldsymbol{\theta}^l$ as

$$D(Q^k||Q^l) = \mathcal{E}_{Q^k}\{l(\mathbf{x};\boldsymbol{\theta}^k) - l(\mathbf{x};\boldsymbol{\theta}^l)\} \tag{31}$$

where $l(\mathbf{x};\boldsymbol{\theta}) = \log p(\mathbf{x};\boldsymbol{\theta})$ and $\mathcal{E}_Q\{f(\cdot)\}$ is a shorthand to $\int_{\mathcal{X}} f(\cdot)p(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}$.

From the preceding analysis, the KLD with respect to the natural parameters can be written as follows

$$D(Q^k||Q^l) = (\boldsymbol{\theta}^k - \boldsymbol{\theta}^l)\cdot\boldsymbol{\eta}^k - (\psi(\boldsymbol{\theta}^k) - \psi(\boldsymbol{\theta}^l)) \tag{32}$$

For a small discrepancy $\delta\boldsymbol{\theta}$, the following approximation holds,

$$D(p(\mathbf{x};\boldsymbol{\theta})||p(\mathbf{x};\boldsymbol{\theta} + \delta\boldsymbol{\theta})) \approx \frac{1}{2}\delta\boldsymbol{\theta}^T G(\theta)\delta\boldsymbol{\theta} \tag{33}$$

which shows that the KLD admits an local interpretation as a quadratic form, with $G(\theta)$ being the Hessian. This justifies its frequent use as the natural metric on the manifold of parametric probability models; it is induced directly by the KLD, which is a reasonable distance to rely on.

The corresponding expression for the dual coordinates is given by

$$\tilde{D}(Q^k||Q^l) = (\boldsymbol{\eta}^k - \boldsymbol{\eta}^l) \cdot \boldsymbol{\theta}^k - (\phi(\boldsymbol{\eta}^k) - \phi(\boldsymbol{\eta}^l)) \quad (34)$$

and it is straightforward to verify that $\tilde{D}(Q^k||Q^l) = D(Q^l||Q^k)$.

The derivatives of the expressions in (32) and (34) with respect to $\boldsymbol{\theta}^k$ can be easily seen to have as follows

$$\nabla_{\boldsymbol{\theta}^k} D(Q^k||Q^l) = G(\boldsymbol{\theta}^k)(\boldsymbol{\theta}^k - \boldsymbol{\theta}^l) \quad (35)$$

and

$$\nabla_{\boldsymbol{\theta}^k} D(Q^l||Q^k) = \boldsymbol{\eta}^k - \boldsymbol{\eta}^l. \quad (36)$$

Using (19) and assuming that the two distributions are close enough, we may approximate (35) by $\boldsymbol{\eta}^k - \boldsymbol{\eta}^l$, like

$$G(\boldsymbol{\theta}^k)(\boldsymbol{\theta}^k - \boldsymbol{\theta}^l) = (\nabla_{\boldsymbol{\theta}} \boldsymbol{\eta}(\boldsymbol{\theta}))\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^k}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^l) \approx \boldsymbol{\eta}^k - \boldsymbol{\eta}^l \quad (37)$$

i.e. to assume linearity for a small area around $\boldsymbol{\theta}^k$. At a first glance, such an approximation seems to be crude. However, is justified by the locality of the mean shift iteration. By considering (12), one may notice that the contribution of each point to the mean shift vector fades out exponentially with their squared distance from the current position. The mean shift vector is dominated by the neighborhood of the current position, making such an approximation possible.

We will investigate the use of symmetrized KLDs, in order to obtain a symmetric kernel. The summation of the two expressions behaves locally (i.e. for $\boldsymbol{\theta}^k$ sufficiently close to $\boldsymbol{\theta}^l$) as the squared distance over the manifold of distributions, [9]. Hence, it seems natural to utilize it in order to define our kernel. The summation approach will be denoted by $D_s(Q^k||Q^l)$ and as we showed, its derivative with respect to $\boldsymbol{\theta}^k$ is approximated by

$$\nabla_{\boldsymbol{\theta}^k} D_s(Q^k||Q^l) = 2(\boldsymbol{\eta}^k - \boldsymbol{\eta}^l) \quad (38)$$

A second symmetric form of the KLD is twice the harmonic mean of (32) and (34), i.e.

$$D_h(Q^k||Q^l) = 4(D(Q^k||Q^l)^{-1} + D(Q^l||Q^k)^{-1})^{-1} \quad (39)$$

The derivative of the above expression has as follows

$$\nabla_{\boldsymbol{\theta}^k} D_h(Q^k||Q^l) = 4C(\boldsymbol{\eta}^k - \boldsymbol{\eta}^l) \quad (40)$$

where

$$C = \frac{D(Q^k||Q^l)^{-2} + D(Q^l||Q^k)^{-2}}{(D(Q^k||Q^l)^{-1} + D(Q^l||Q^k)^{-1})^2} \quad (41)$$

A rationale for the harmonic mean can be found in [15], along with many other interesting approaches and distances.

To summarize the section, what we have shown is that the interplay between the natural and the expectation parameters can be very beneficial in order to define the derivatives of the KLD-derived distances.

4.2. An kernel based on the entropic prior

Based on the above analysis, a probabilistic kernel can be derived. Our kernel has the following form

$$K(Q^*; Q^k) \propto q_k^P \exp\left(-q_k D(Q^*||Q^k)\right) \sqrt{|G(\boldsymbol{\theta}^*)|} \quad (42)$$

The above kernel is a conjugate prior for $\boldsymbol{\theta}^*$ centered at $\boldsymbol{\theta}^k$, (see [16] for a delicate derivation) and q_k a function of the sample size m^k. To be more precise, it corresponds to the Normal-Wishart prior, i.e. a common prior used for the multivariate normal distribution, when mean and covariance are both unknown. Furthermore, q_i is a parameter to encode the balance between the degree of confidence in $\boldsymbol{\theta}^i$ against the uninformative Jeffreys prior $\propto \sqrt{|G(\boldsymbol{\theta}^*)|}$. This trade-off corresponds to the smoothing parameter of the original mean shift and we may place $q_k = \frac{m^k}{\sigma^2}$. However, a straightforward use of the sample size based on the Cramer-Rao bound is not effective when dealing with models that are highly misspecified, i.e. when the speaker-model that generates the data does not belong to the family of distributions we deploy. Despite the fact that the use of single Gaussians leads to very fast algorithms, is remains a highly misspecified model to describe the multimodal data generating process of the utterances of a speaker. Therefore, in this paper we will adopt a more simple and heuristic approach and use unnormalized kernels (as in [17]) without involving the sample sizes. Hence, we will ignore both $\sqrt{|G(\boldsymbol{\theta}^*)|}$ and q_k in (42), knowing that any approach that does not involve the sample size (i.e. the variance when estimating $\{\theta^k\}_{k=1}^n$) is clearly suboptimal for the speaker clustering task.

Note also that the formula in (42) may be used for mixtures of Gaussians, too. However, this holds only if the complete-data likelihood $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ is considered, where $\mathbf{z}$ are the component indicators of the observations $\mathbf{x}$, [16]. The reason is that while the marginal density $p(\mathbf{x}|\boldsymbol{\theta})$ does not belong to the exponential family, the density of the complete data $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ does, and therefore it meets the demands for having a conjugate prior. Although $\mathbf{z}$ are unknown, their MAP-estimate can easily be obtained by the final E-step of the EM algorithm.

We further note that its dual kernel (i.e. with its dual KLD $D(Q^k||Q^*)$ into the exponent) may also be considered, since it corresponds to the Normal-Inverse Wishart prior, [12]. Moreover, we also try to place in our kernel the two symmetrized versions of the KLD, discussed above. The analysis remains the same if the approximation discussed in (37) is adopted. For a discussion regarding the priors that are obtained by symmetric versions of the KLD we refer to [18].

Finally, we choose to differentiate the kernels with respect to the natural parameters, so that the new position of the distribution is calculated by averaging the expectation parameters. Recall that averaging in the $\boldsymbol{\eta}$-coordinates is closer to our task, in the following sense

$$\hat{\boldsymbol{\eta}} = \frac{m^k}{m^k + m^l}\boldsymbol{\eta}^k + \frac{m^l}{m^k + m^l}\boldsymbol{\eta}^l, \quad (43)$$

i.e. is compatible to the closed-form expression we use to obtain the ML estimate when merging two (or more) clusters in the hierarchical clustering algorithm. We finally note that the approach of averaging in the $\boldsymbol{\theta}$-coordinates has also be examined, but the performance of the algorithm showed a slight degradation.

5. The proposed algorithm

Having covered much of the theoretical background of the mean shift algorithm as well as some of the main properties of the exponential family, we now describe the method we propose. Suppose we are given a collection of K speech segments, and let $\mathbf{\Theta} = (\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^K)$ be an estimate of their natural parameters. We consider here only maximum likelihood (ML) estimates, however other types of estimators can be applied as well (MAP, M-estimators, etc.). We also denote by $\mathbf{H} = (\boldsymbol{\eta}^1, \boldsymbol{\eta}^2, \ldots, \boldsymbol{\eta}^K)$ and $\mathbf{m} = (m^1, m^2, \ldots, m^K)$ the corresponding expectation parameters and sample sizes, respectively.

5.1. Derivation of the mean shift iteration

In order to derive the iteration of the mean shift, we should first consider the density parametrized by $\boldsymbol{\theta}$, as the posterior density, given the observations $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$ and an initial labeling $\mathbf{Z} = (z^{(1)}, \ldots, z^{(n)})$ that corresponds to the initial segmentation. The proposed expression is as follows

$$\pi(\boldsymbol{\theta}^*|\mathbf{\Theta}) \propto \frac{1}{K}\sum_{k=1}^{K} \exp\left(-h_s^{-2} D_{\cdot}(Q^*||Q^k)\right) \qquad (44)$$

where Q^k denotes the density parametrized by $\boldsymbol{\theta}^k$ or equivalently $\boldsymbol{\eta}^k$. The above expression may be regarded as the posterior of $\boldsymbol{\theta}$, given $(\mathbf{X}, \mathbf{Z})$. We use $h_s > 0$ as the tuning parameter, common to all segments, which as explained in Sect. 4.2 is clearly a suboptimal approach, since it should be encoding information about the sample sizes $\mathbf{m}$. From the analysis in Sect. 4, by differentiating (44) with respect to $\boldsymbol{\theta}$, the mean shift vector vanishes when

$$\boldsymbol{\eta}^* = \frac{\sum_{k=1}^{K} \boldsymbol{\eta}^k p_k \exp\left(-h_s^{-2} D_{\cdot}(Q^*||Q^k)\right)}{\sum_{k=1}^{K} p_k \exp\left(-h_s^{-2} D_{\cdot}(Q^*||Q^k)\right)}. \qquad (45)$$

In the above expression, the dot subscript is placed to denote which KLD will be used. The terms $\{p_k\}_{k=1}^{K}$ are placed in order to attach information about the proximity of the segments in the time domain. Time domain (or temporal information) has no meaning unless the segments are parts of a unique audio file, i.e. when the algorithm operates in the speaker diarization domain. The use of temporal information is discussed in Sect. 5.2.

To express the result in (45) in terms of the mean shift vector, we obtain

$$\mathbf{m}_{h_s,h_t}(\boldsymbol{\eta}_t) = \frac{\sum_{k=1}^{K} \boldsymbol{\eta}^k p_k \exp\left(-h_s^{-2} D_{\cdot}(Q_t||Q^k)\right)}{\sum_{k=1}^{K} p_k \exp\left(-h_s^{-2} D_{\cdot}(Q_t||Q^k)\right)} - \boldsymbol{\eta}_t. \qquad (46)$$

Similarly to the original algorithm described in Sect. 2.3, the proposed algorithm has the following form.

For each segment $k = 1, 2 \ldots, K$ set $t = 0$, $\boldsymbol{\eta}_t \leftarrow \boldsymbol{\eta}^k$

1. calculate $\mathbf{m}_{h_s,h_t}(\boldsymbol{\eta}_t)$
2. set $\boldsymbol{\eta}_{t+1} \leftarrow \boldsymbol{\eta}_t + \mathbf{m}_{h_s,h_t}(\boldsymbol{\eta}_t)$
3. if $D_{\cdot}(Q_t||Q_{t+1}) < \epsilon$ goto 4; else $t \leftarrow t+1$ and goto 1.
4. store $\tilde{\boldsymbol{\eta}}^k = \boldsymbol{\eta}_{t+1}$.

where (h_s, h_t) denote the bandwidths on the spectral and temporal domain, respectively. The matrix $\tilde{\mathbf{H}} = (\tilde{\eta}^1, \tilde{\boldsymbol{\eta}}^2, \ldots, \tilde{\boldsymbol{\eta}}^K)$ now holds the convergent points of $\mathbf{H}$. Note that the conversion from $\boldsymbol{\eta}$ to $\boldsymbol{\theta}$ and vice versa is unnecessary; all the calculations lie on the $\boldsymbol{\eta}$-parametrization.

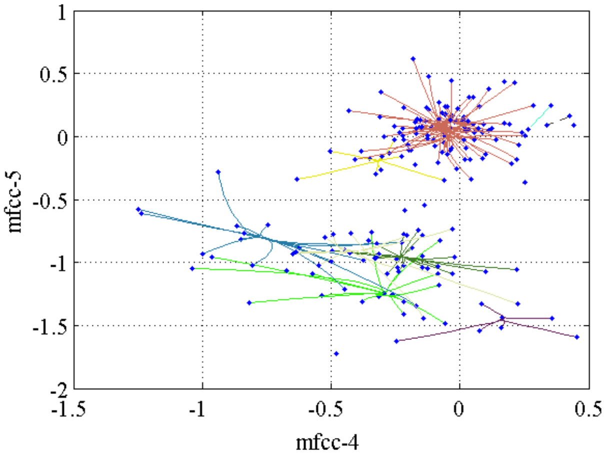

Figure 1: *Illustration of the convergence with real data. xy-axes: mean value of 4-th and 5-th mfcc coefficient. Blue dots correspond to the initial position of the segments. Trajectories that were attracted by the same mode are depicted with the same color. 191 segments merged into 16 clusters. 6 clusters are singletons.*

5.2. Making use of the temporal information

As mentioned above, one can make use of the temporal information in order to enhance the results and avoid having abrupt changes into the derived clustering. The proximity of the entities is used in the original mean shift as well. The kernel is multiplied by a spatial kernel that is a function of the distance between the pixel in question of the others. Color segmentation algorithms result to more smoothed images when they operate on the joint spatial-color range domain. In speaker diarization, many algorithms are based on an HMM framework, where the self-transition probability is set to be orders of magnitude higher that the probability of moving to other states.

In the proposed method, we have derived a simple yet effective way to make use of this information. Let us denote by $\mathbf{t} = (t^1, t^2, \ldots, t^n)$ the central value of their time index, i.e $t^i = \frac{t_s^i + t_e^i}{2}$, where t_s^i and t_e^i denote the first and last time index of the ith speech segment. Let also introduce a temporal kernel $k_t(\cdot)$ using the Cauchy density

$$k_t(t^k, t^l) = \frac{1}{\pi}\frac{h_t}{(t^k - t^l)^2 + h_t^2}. \qquad (47)$$

We choose the Cauchy density since it has much heavier tails when compared to the Gaussian one. A Gaussian temporal kernel would make the contribution of segments lying far from the target one in the temporal domain exponentially small, which is an undesired property. Note also that as (45) shows, we do not differentiate this kernel, since we do not attempt to obtain modes in the joint speaker-temporal domain but rather to incorporate into the model our prior knowledge of temporal continuity of the speaker labels. To conclude, we suggest the use of the above temporal kernel, especially when dealing with short speech segments in the speaker diarization domain. To do so, we set $p_k = \frac{1}{\pi}\frac{h_t}{(t^k - t^*)^2 + h_t^2}$, where t^* is the central time index of the target segment and h_t denotes the bandwidth in the temporal domain.

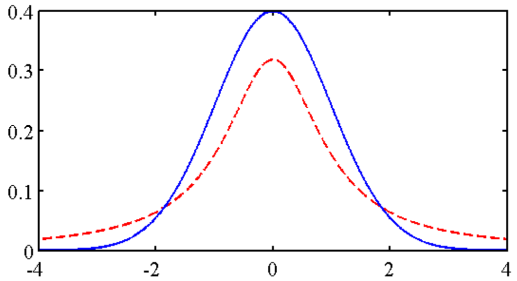

Figure 2: *Cauchy density (red dashed line) vs. Gaussian density (blue solid line). Note the heavy tails of the Cauchy density*

6. Experiments

6.1. Set-up and datasets

We tested our algorithm using the ESTER Speaker Diarization benchmark, [19]. ESTER is a very rich Broadcast News (BN) corpus, consisting of 32 shows from various France Radio Channels. The shows are divided into the development (14 shows, about 8 hours total duration, denoted by ESTER-DEV) and the test set (18 shows, about 10 hours total duration, denoted by ESTER-TEST). To compare our algorithm with the hierarchical clustering approach, we used the open source software provided by the LIUM Laboratory, [20], where the local-ΔBIC is deployed as the dissimilarity measure.

As explained above, both algorithms are based on the step-by-step approach to speaker diarization, i.e. they operate on the speaker clustering stage. To do so, the standard segmentation technique (i.e. a sliding window) is first applied to the mfcc stream, using the LIUM software. As front-end features, we used 18-dimensional static mfcc, augmented by the log-energy, while no Viterbi re-alignment is applied.

Furthermore, the co-called linear clustering stage, i.e. the merging of the consecutive segments prior to the main clustering stage, in order to obtain longer segments has been applied only to the hierarchical clustering. We did so, since we wanted to test our approach without resorting to methods that merge segments in an explicit way. The mean shift algorithm merges the segments only when the modes of the underlying pdf have been detected. Hence, methods such as the linear clustering are out of the scope of the mean shift algorithm. As an alternative to the linear clustering step, the approach of utilizing the temporal information via the Cauchy density is examined, since is much closer to the spirit of the mean shift algorithm.

6.2. Experimental results

The first experiment was carried out using the ESTER-DEV set. The best performance attained by the competing methods is illustrated in Fig. 4. Based on the optimal parameter values of the development set, we ran the same algorithm for the ESTER-TEST set, and the results are depicted in Fig. 5.

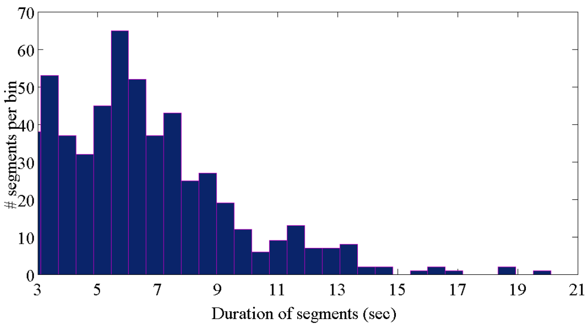

Figure 3: *Typical histogram of the duration of the segments, without applying the linear clustering algorithm. Note that about half of the segment having duration < 5 correspond to non-speech segments.*

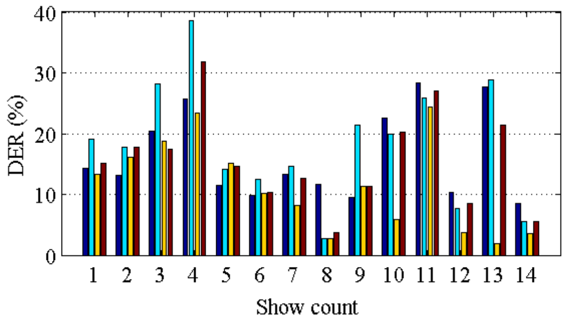

Figure 4: *Best performance in terms of diarization error rates for the ESTER-DEV set. From left to right: HC with local ΔBIC, MS with summed KLD, MS with harmonic-mean symmetrized KLD, MS with asymmetric KLD.*

The results - summarized in Table 1 - show that the performance attained by the proposed method is comparable to the standard paradigm of the BIC-based HC. We should emphasize though that the strength of the mean shift algorithm compared to the HC can be estimated by considering the fact that when the KLD-harmonic mean is deployed instead of the ΔBIC, the performance of the HC degrades to DER $> 30\%$. We should further notice that many aspects remain open, such as the use of the sample size into the KLD, that may lead to much better results. Moreover, several other techniques, such as the variable bandwidth should also be considered, [14]. The literature of non-parametric estimation is rich-enough to provide us with many such techniques. Finally, different divergences should also be examined, such as the Hellinger and other members of the family of f-divergence, which may be more appropriate for the specific task.

7. Discussion and technicalities

7.1. Speeding-up the algorithm

The complexity of the algorithm is $\mathcal{O}(\bar{t}K^2)$, i.e. of the same order with the hierarchical clustering multiplied by the average number of iterations, $\bar{t}$. However, the algorithm can be accelerated by exploiting its locality with respect to the target segment. Using the fact that the convergent point of the target segment cannot be far away from its initial position, one may prune the algorithm by setting a lower bound to the KLDs. This can be applied right after the first iteration of each target segment. After some experiments with broadcast news datasets, we concluded that nearly 80% of the segments can be discarded without affecting the performance of the algorithm. This pruning technique is highly suggested, especially when dealing with audio files of 30 min duration and above. Note also that it corresponds approximately to the use of a truncated version of the Gaussian

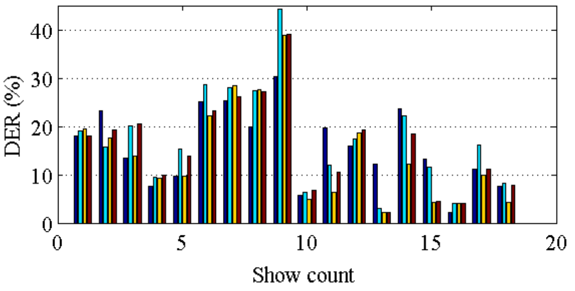

Figure 5: *Diarization error rates for the ESTER-TEST set. From left to right: HC with Local–ΔBIC, MS with summed KLD, MS with harmonic-mean symmetrized KLD, MS with asymmetric KLD. The parameters have been optimized based on the ESTER-DEV set.*

Table 1: Overall Speaker Diarization Error Rate (%) on ESTER

	ESTER-DEV	ESTER-TEST
HC Local-BIC	15.76	16.28
MS summed KLD	18.78	17.77
MS Harmonic mean KLD	14.88	15.49
MS asymmetric KLD	16.55	16.83
False Alarm Rate	0.3	0.6
Missed Speech Rate	0.9	1.2

kernel, which retains the main properties of the non-truncated version, [7]. Other pruning approaches can be found in [21].

Finally, note that the algorithm admits *parallel processing* solutions; the convergent point of each segment is independent of the convergent points of the other segments. Hence, the computation of the convergent points can be distributed to many processors without affecting the final results. This is in contrast to the hierarchical clustering and EM-based algorithms, where the outcome of the k-th iteration is required as input to the $(k+1)$-th iteration.

7.2. Merging the convergent points into clusters

As mentioned in Sec. 2, the convergence of two or more distinct entities to exactly the same point (i.e. mode) is an unrealistic demand, especially when a temporal kernel is deployed. A certain amount of tolerance ϵ of the discrepancy between the convergent point should be introduced. Hence, a further algorithmic step should also be appended, in order to form the final clusters based on the convergent points. One may use the standard hierarchical clustering, however faster approaches can be applied. Consider the following procedure.

Let Y be the desired vector of the cluster labels. We use C denote the current number of clusters. Set $Y(1) = 1$ and $C = 1$ and let $\bar{\boldsymbol{\eta}}^1 = \tilde{\boldsymbol{\eta}}^1$ denote the expectation parameters of the first cluster. For the remaining convergent points $\tilde{\boldsymbol{\eta}_k}, k = 2, 3 \ldots, K$ do

```
1. for $c = 1 : C$
    • $d(k,c) = D_s(\tilde{Q}^k || \bar{Q}^c)$
    • if $\min_c d(k,c) \leq \epsilon$, $c^* = \text{argmin}_c\, d(k,c)$,
      $\bar{\boldsymbol{\eta}}^{c*} \leftarrow \text{merge}(\bar{\boldsymbol{\eta}}^{c*}, \tilde{\boldsymbol{\eta}}^k), Y(k) = c^*$
    • else
      $C \leftarrow C + 1, \bar{\boldsymbol{\eta}}^C \leftarrow \tilde{\boldsymbol{\eta}}^i, Y(k) = C$
```

where the merging is carried out like

$$\bar{\boldsymbol{\eta}}^{c*} \leftarrow \frac{m^{c*}}{m^{c*} + m^k}\bar{\boldsymbol{\eta}}^{c*} + \frac{m^k}{m^{c*} + m^k}\tilde{\boldsymbol{\eta}}^k, \tag{48}$$

i.e. the usual weighted average. With the above "linearized" clustering algorithm, we end-up having the number of clusters C, along with the assignments Y. After several experiments, we concluded that the above fast clustering procedure yields almost identical results to the hierarchical clustering.

In order to avoid the use of a new threshold and adjustable parameters, we use the (local) ΔBIC as the distance, with no tuning parameter (i.e. $\lambda = 1$, its theoretically correct value). Note that if $\lambda = 1$ were used to performed clustering directly with the initial values of $\{\boldsymbol{\eta}_k\}_{k=1}^K$ (instead of their convergent points $\{\tilde{\boldsymbol{\eta}}_k\}_{k=1}^K$) very few segments would merge. As such, it should be considered as a reference parameter-free test of similarity, ensuring us that the backbone of the overall proposed method is indeed the mean shift algorithm and not this final step.

8. Conclusion and future work

In this paper, we introduced the mean shift algorithm to the problem of speaker clustering. The proposed algorithm should be consider as an alternative to the hierarchical clustering approach, which remains the baseline technique, at least when a point estimate of the partition is required. By restricting ourselves to the exponential family, we derived some necessary maths, in order to make the algorithm capable of operating on the family of parametric models. Several technical aspect, concerning the appropriate distance measures, as well as the use of the temporal information have also been discussed.

The algorithm is a new framework to the problem of speaker clustering. Hence, alternative formulations can be derived, such as novel front-end features, as well as models derived by the combination of the GMM/UBM model with dimensionality reduction techniques (UBM-supervectors, eigenvoices, fishervoices, etc., [22], [23]). Furthermore, a deeper analysis of the parameter space as a Riemannian manifold via Information Geometry may lead to more appropriate distances than the simple KLD that we utilized, and enhance the performance of the algorithm.

9. Acknowledgements

This work is funded by the Greek General Secretariat of Research and Technology under the program PENED-03/251.

10. References

[1] D.A. Reynolds P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of ICASSP*, 2005, pp. V–953–V 956.

[2] S. Meignier et al., "Step-by-Step and Integrated Approaches in Broadcast News Speaker Diarization," *Elsevier Computer Speech and Language*, pp. 303 – 330, April-July 2006.

[3] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky, "The Sticky HDP-HMM: Bayesian nonparametric Hidden Markov Models with Persistent States," 2009.

[4] Fabio Valente, *Variational Bayesian methods for audio indexing*, Ph.D. thesis, September 2005.

[5] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 1557–1565, 2006.

[6] Changjiang Yang, Ramani Duraiswami, and Larry Davis, "Efficient mean-shift tracking via a new similarity measure," in *IEEE Computer Society*, 2005, pp. 176–183.

[7] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790 – 799, August 1995.

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

[9] Robert E. Kass, "The Geometry of Asymptotic Inference," *Statistical Science*, vol. 4, no. 3, pp. 188–219, August 1989.

[10] S. Yoshizawa and K. Tanabe, "Dual differential geometry associated with kullback-leibler information on the Gaussian distributions and its 2-parameter deformations," *SUT Journal of Mathematics*, vol. 25, no. 1, pp. 113–137, 1999.

[11] Frank Nielsen and Richard Nock, "The entropic centers of multivariate normal distributions," in *European Workshop on Computational Geometry (EuroCG)*, Nancy, France, March 2008, pp. 221–224.

[12] Hichem Snoussi, "The geometry of prior selection," *Neurocomputing*, vol. 67, pp. 214–244, 2005.

[13] 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.

[14] 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.

[15] Don H. Johnson and Sinan Sinanovic, "Symmetrizing the Kullback-Leibler Distance," Tech. Rep., IEEE Transactions on Information Theory, 2000.

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

[17] Raghav Sabbarao, *Robust Statistics over Riemannian Manifolds for Computer Vision*, Ph.D. thesis, Rutgers, N.J., May 2008.

[18] M. J. Bayarri and G. Garcia-Donato, "Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing," *Journal Of The Royal Statistical Society Series B*, vol. 70, no. 5, pp. 981–1003, 2008.

[19] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER evaluation campaign for the Rich Transcription of French Broadcast News," in *Proc. Language Evaluation and Resources Conference*, 2006.

[20] 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.

[21] B. Georgescu, I. Shimshoni, and P. Meer, "Mean shift based clustering in high dimensions: a texture classification example," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2003, pp. 456–463 vol.1.

[22] R. Kuhn, "Speaker verification and speaker identification based on eigenvoices ," *Acoustical Society of America Journal*, vol. 109, June 2001.

[23] W. Tsai, S. Cheng, Y. Chao, and H. Wang, "Clustering speech utterances by the speaker using eigenvoice-motivated vector space models," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Philadelphia, USA*, March 2005, pp. 725–728.