Clustering Multivariate Normal Distributions

Frank Nielsen[1,2] and Richard Nock[3]

[1] LIX — Ecole Polytechnique, Palaiseau, France
nielsen@lix.polytechnique.fr
[2] Sony Computer Science Laboratories Inc., Tokyo, Japan
nielsen@csl.sony.co.jp
[3] Ceregmia — Université Antilles-Guyane, Schoelcher, France
rnock@martinique.univ-ag.fr

Abstract. In this paper, we consider the task of clustering multivariate normal distributions with respect to the relative entropy into a prescribed number, k, of clusters using a generalization of Lloyd's k-means algorithm [1]. We revisit this information-theoretic clustering problem under the auspices of mixed-type Bregman divergences, and show that the approach of Davis and Dhillon [2] (NIPS*06) can also be derived directly, by applying the Bregman k-means algorithm, once the proper vector/matrix Legendre transformations are defined. We further explain the dualistic structure of the sided k-means clustering, and present a novel k-means algorithm for clustering with respect to the symmetrical relative entropy, the J-divergence. Our approach extends to differential entropic clustering of arbitrary members of the same exponential families in statistics.

1 Introduction

In this paper, we consider the problem of clustering multivariate normal distributions into a given number of clusters. This clustering problem occurs in many real-world settings where each datum point is naturally represented by multiple observation samples defining a mean and a variance-covariance matrix modeling the underlying distribution: Namely, a multivariate normal (Gaussian) distribution. This setting allows one to conveniently deal with anisotropic noisy data sets, where each point is characterized by an individual Gaussian distribution representing locally the amount of noise. Clustering "raw" normal data sets is also an important algorithmic issue in computer vision and sound processing. For example, Myrvoll and Soong [3] consider this task for adapting hidden Markov model (HMM) parameters in a structured maximum a posteriori linear regression (SMAPLR), and obtained improved speech recognition rate. In computer vision, Gaussian mixture models (GMMs) abound from statistical image modeling learnt by the expectation-maximization (EM) soft clustering technique [4], and therefore represent a versatile source of raw Gaussian data sets to manipulate efficiently. The closest prior work to this paper is the differential entropic clustering of multivariate Gaussians of Davis and Dhillon [2], that can be derived from our framework as a special case.

F. Nielsen (Ed.): ETVC 2008, LNCS 5416, pp. 164–174, 2009.

A central question for clustering is to define the appropriate information-theoretic measure between any pair of multivariate normal distribution objects as the Euclidean distance falls short in that context. Let $N(m, S)$ denote[1] the d-variate normal distribution with mean m and variance-covariance matrix S. Its probability density function (pdf.) is given as follows [5]:

$$p(x; m, S) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det S}} \exp\left(-\frac{(x-m)^T S^{-1}(x-m)}{2}\right), \quad (1)$$

where $m \in \mathbb{R}^d$ is called the mean, and $S \succ 0$ is a positive semi-definite matrix called the variance-covariance matrix, satisfying $x^T S x \geq 0 \ \forall x \in \mathbb{R}^d$. The variance-covariance matrix $S = [S_{i,j}]_{i,j}$ with $S_{i,j} = E[(X^{(i)} - m^{(i)})(X^{(j)} - m^{(j)})]$ and $m_i = E[X^{(i)}] \ \forall i \in \{1, ..., d\}$, is an invertible symmetric matrix with positive determinant: $\det S > 0$. A normal distribution "statistical object" can thus be interpreted as a "compound point" $\tilde{\Lambda} = (m, S)$ in $D = \frac{d(d+3)}{2}$ dimensions by stacking the mean vector m with the $\frac{d(d+1)}{2}$ coefficients of the symmetric variance-covariance matrix S. This encoding may be interpreted as a serialization or linearization operation. A fundamental distance between statistical distributions that finds deep roots in information theory [6] is the *relative entropy*, also called the Kullback-Leibler divergence or information discrimination measure. The oriented distance is asymmetric (ie., $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$) and defined as:

$$\mathrm{KL}(p(x; m_i, S_i)||p(x; m_j, S_j)) = \int_{x \in \mathbb{R}^d} p(x; m_i, S_i) \log \frac{p(x; m_i, S_i)}{p(x; m_j, S_j)} \mathrm{d}x. \quad (2)$$

The Kullback-Leibler divergence expresses the *differential relative entropy* with the *cross-entropy* as follows:

$$\mathrm{KL}(p(x; m_i, S_i)p(x; m_j, S_j)) = -H(p(x; m_i, S_i)) - \int_{x \in \mathbb{R}^d} p(x; m_i, S_i) \log p(x; m_j, S_j) \mathrm{d}x, \quad (3)$$

where the Shannon' differential entropy is

$$H(p(x; m_i, S_i)) = -\int_{x \in \mathbb{R}^d} p(x; m_i, S_i) \log p(x; m_i, S_i) \mathrm{d}x, \quad (4)$$

independent of the mean vector:

$$H(p(x; m_i, S_i)) = \frac{d}{2} + \frac{1}{2} \log(2\pi)^d \det S_i. \quad (5)$$

Fastidious integral computations yield the well-known Kullback-Leibler divergence formula for multivariate normal distributions:

[1] We do not use the conventional (μ, Σ) notations to avoid misleading formula later on, such as $\sum_{i=1}^{n} \Sigma_i$, etc.

$$\mathrm{KL}(p(x;m_i,S_i)||p(x;m_j,S_j)) = \frac{1}{2}\log|S_i^{-1}S_j| + \frac{1}{2}\mathrm{tr}\left((S_i^{-1}S_j)^{-1}\right) - \frac{d}{2} + \frac{1}{2}(m_i-m_j)^T S_j^{-1}(m_i-m_j), \quad (6)$$

where $\mathrm{tr}(S)$ is the trace of square matrix S, the sum of its diagonal elements: $\mathrm{tr}(S) = \sum_{i=1}^{d} S_{i,i}$. In particular, the Kullback-Leibler divergence of normal distributions reduces to the quadratic distance for unit spherical Gaussians: $\mathrm{KL}(p(x;m_i,I)||p(x;m_j,I)) = \frac{1}{2}||m_i - m_j||^2$, where I denotes the $d \times d$ identity matrix.

2 Viewing Kullback-Leibler Divergence as a Mixed-Type Bregman Divergence

It turns out that a neat generalization of both statistical distributions and information-theoretic divergences brings a *simple way* to find out the same result of Eq. 6 by *bypassing* the integral computation. Indeed, the well-known normal density function can be expressed into the canonical form of *exponential families* in statistics [7]. Exponential families include many familiar distributions such as Poisson, Bernoulli, Beta, Gamma, and normal distributions. Yet exponential families do not cover the full spectrum of usual distributions either, as they do not contain the uniform nor Cauchy distributions.

Let us first consider univariate normal distributions $N(m, s^2)$ with associated probability density function:

$$p(x;m,s^2) = \frac{1}{s\sqrt{2\pi}}\exp - \left(\frac{(x-m)^2}{2s^2}\right). \quad (7)$$

The pdf can be mathematically rewritten to fit the *canonical decomposition* of distributions belonging to the exponential families [7], as follows:

$$p(x;m,s^2) = p(x;\theta=(\theta_1,\theta_2)) = \exp\left\{ < \theta, t(x) > -F(\theta) + C(x)\right\}, \quad (8)$$

where $\theta = (\theta_1 = \frac{\mu}{\sigma^2}, \theta_2 = -\frac{1}{2\sigma^2})$ are the *natural parameters* associated with the *sufficient statistics* $t(x) = (x, x^2)$. The *log normalizer* $F(\theta) = -\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log\frac{-\pi}{\theta_2}$ is a strictly convex and differentiable function that specifies uniquely the exponential family, and the function $C(x)$ is the carrier measure. See [7,8] for more details and plenty of examples. Once this canonical decomposition is figured out, we can simply apply the generic *equivalence theorem* [9] [8] Kullback-Leibler↔Bregman divergence [10]:

$$\mathrm{KL}(p(x;m_i,S_i)||p(x;m_j,S_j)) = D_F(\theta_j||\theta_i), \quad (9)$$

to get the closed-form formula easily. In other words, this theorem (see [8] for a proof) states that the Kullback-Leibler divergence of two distributions of the *same* exponential family is equivalent to the Bregman divergence for the log normalizer generator by *swapping* arguments. The Bregman divergence [10] D_F is defined as the tail of a Taylor expansion for a strictly convex and differentiable function F as:

$$D_F(\theta_j||\theta_i) = F(\theta_j) - F(\theta_i) - < \theta_j - \theta_i, \nabla F(\theta_i) >, \quad (10)$$

where $< \cdot, \cdot >$ denote the vector inner product ($< p, q >= p^T q$) and ∇F is the gradient operator. For multivariate normals, the same kind of decomposition exists but on *mixed-type* vector/matrix parameters, as we shall describe next.

3 Clustering with Respect to the Kullback-Leibler Divergence

3.1 Bregman/Kullback-Leibler Hard k-Means

Banerjee et al. [9] generalized Lloyd's k-means hard clustering technique [1] to the broad family of Bregman divergences D_F. The Bregman hard k-means clustering of a point set $\mathcal{P} = \{p_1, ..., p_n\}$ works as follows:

1. **Initialization.** Let $C_1, ..., C_k$ be the initial k cluster centers called the seeds. Seeds can be initialized in many various ways and is an important step to consider in practice, as explained in [11]. The simplest technique, called Forgy's initialization [12], is to allocate *at random* seeds from the source points.
2. **Repeat until converge or stopping criterion is met**
   (a) **Assignment.** Associate to each "point" p_i its closest center with respect to divergence D_F: $p_i \rightarrow \arg\min_{C_j \in \{C_1,...,C_k\}} D_F(p_i||C_j)$. Let $\mathcal{C}_l$ denote the lth cluster, the set of points closer to center C_l than to any other cluster center. The clusters form a partition of the point set $\mathcal{P}$. This partition may be geometrically interpreted as the underlying partition emanating from the Bregman Voronoi diagram of the cluster centers $C_1, ..., C_k$ themselves, see [8].
   (b) **Center re-estimation.** Choose the new cluster centers $C_i \; \forall i \in \{1, ..., k\}$ as the cluster respective centroids: $C_i = \frac{1}{|\mathcal{C}_i|} \sum_{p_j \in \mathcal{C}_i} p_j$. A key property emphasized in [9] is that the Bregman centroid defined as the minimizer of the right-side intracluster average $\arg\min_{c \in \mathbb{R}^d} \sum_{p_i \in \mathcal{C}_l} D_F(p_i||c)$ is *independent* of the considered Bregman generator F, and always coincide with the center of mass.

The Bregman hard clustering enjoys the same convergence property as the traditional k-means. That is, the *Bregman loss* function $\sum_{l=1}^{k} \sum_{p_i \in \mathcal{C}_l} D_F(p_i||C_l)$ monotonically decreases until convergence is reached. Thus a stopping criterion can also be choosen to terminate the loop as soon as the difference between the Bregman losses of two successive iterations goes below a prescribed threshold. In fact, Lloyd's algorithm [1] is a Bregman hard clustering for the quadratic Bregman divergence ($F(x) = \sum_{i=1}^{d} x_i^2$) with associated (Bregman) quadratic loss. As mentioned above, the centers of clusters are found as right-type sum average minimization problems. For a n-point set $\mathcal{P} = \{p_1, ..., p_n\}$, the center is defined as

$$\arg\min_{c \in \mathbb{R}^d} D_F(p_i||c) = \frac{1}{n} \sum_{i=1}^{n} p_i. \tag{11}$$

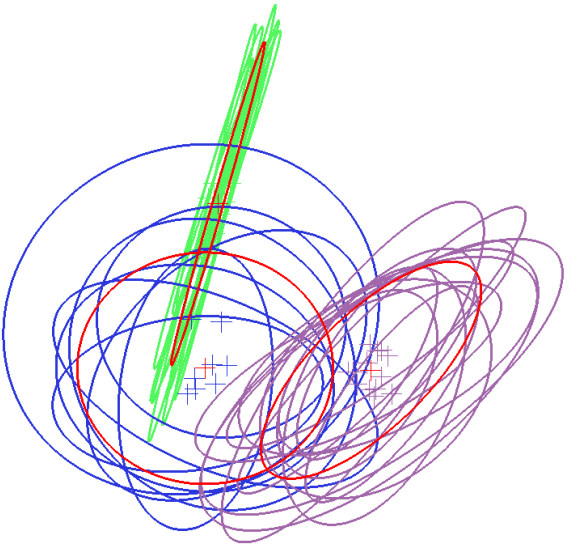


Fig. 1. Bivariate normal k-means clustering ($k = 3$, $d = 2$, $D = 5$) with respect to the right-type Bregman centroid (the center of mass of natural parameters, equivalent to the left-type Kullback-Leibler centroid) of 32 bivariate normals. Each cluster is displayed with its own color, and the centroids are rasterized as red variance-covariance ellipses centered on their means.

That is, the Bregman right-centroid is surprisingly invariant to the considered Bregman divergence [9] and always equal to the center of mass. Note that although the squared Euclidean distance is a Bregman (symmetric) divergence, it is not the case for the single Euclidean distance for which the minimum average distance optimization problem yields the Fermat-Weber point [13] that *does not admit closed-form* solution.

Thus for clustering normals with respect to the Kullback-Leibler divergence using this Bregman hard clustering, we need to consider the oriented distance $D_F(\theta_i||\omega_l)$ for the log normalizer of the normal distributions interpreted as members of a given exponential family, where ω_l denote the cluster centroid in the natural parameter space. Since $D_F(\theta_i||\omega_l) = \mathrm{KL}(c_l||p_i)$ it turns out that the hard Bregman clustering minimizes the Kullback-Leibler loss $\sum_{l=1}^{k} \sum_{p_i \in \mathcal{C}_l} \mathrm{KL}(c_l||p_i)$. We now describe the primitives required to apply the Bregman k-means clustering to the case of the Kullback-Leibler clustering of multivariate normal distributions.

3.2 Mixed-Type Parameters of Multivariate Normals

The density function of multivariate normals of Eq. 1 can be rewritten into the canonical decomposition of Eq. 8 to yield an exponential family of order $D = \frac{d(d+3)}{2}$ (the mean vector and the positive definite matrix S^{-1} accounting

respectively for d and $\frac{d(d+1)}{2}$ parameters). The sufficient statistics is *stacked* onto a two-part D-dimensional vector/matrix entity

$$\tilde{x} = (x, -\frac{1}{2}xx^T) \tag{12}$$

associated with the natural parameter

$$\tilde{\Theta} = (\theta, \Theta) = (S^{-1}m, \frac{1}{2}S^{-1}). \tag{13}$$

Accordingly, the source parameter are denoted by $\tilde{\Lambda} = (m, S)$. The log normalizer specifying the exponential family is (see [14]):

$$F(\tilde{\Theta}) = \frac{1}{4}\mathrm{Tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta + \frac{d}{2}\log 2\pi. \tag{14}$$

To compute the Kullback-Leibler divergence of two normal distributions $N_p = \mathcal{N}(\mu_p, \Sigma_p)$ and $N_q = \mathcal{N}(\mu_q, \Sigma_q)$, we use the Bregman divergence as follows:

$$\mathrm{KL}(N_p||N_q) = D_F(\tilde{\Theta}_q||\tilde{\Theta}_p) \tag{15}$$

$$= F(\tilde{\Theta}_q) - F(\tilde{\Theta}_p) - < (\tilde{\Theta}_q - \tilde{\Theta}_p), \nabla F(\tilde{\Theta}_p) > . \tag{16}$$

The inner product $< \tilde{\Theta}_p, \tilde{\Theta}_q >$ is a *composite* inner product obtained as the sum of two inner products of vectors and matrices:

$$< \tilde{\Theta}_p, \tilde{\Theta}_q > = < \Theta_p, \Theta_q > + < \theta_p, \theta_q > . \tag{17}$$

For matrices, the inner product $< \Theta_p, \Theta_q >$ is defined by the trace of the matrix product $\Theta_p\Theta_q^T$:

$$< \Theta_p, \Theta_q > = \mathrm{Tr}(\Theta_p\Theta_q^T). \tag{18}$$

Figure 1 displays the Bregman k-means clustering result on a set of 32 bivariate normals.

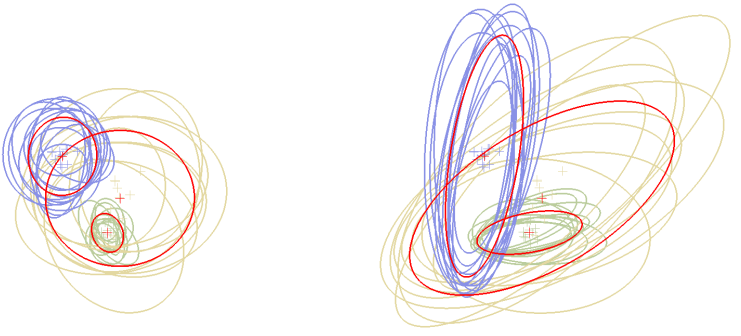
4 Dual Bregman Divergence

We introduce the Legendre transformation to interpret *dually* the former k-means Bregman clustering. We refer to [8] for detailed explanations that we concisely summarize here as follows: Any Bregman generator function F admits a *dual* Bregman generator function $G = F^*$ via the Legendre transformation

$$G(y) = \sup_{x \in \mathcal{X}} \{< y, x > -F(x)\}. \tag{19}$$

The supremum is reached at the *unique* point where the gradient of $G(x) = < y, x > -F(x)$ vanishes, that is when $y = \nabla F(x)$. Writing $\mathcal{X}'_F$ for the *gradient space* $\{x' = \nabla F(x) | x \in \mathcal{X}\}$, the convex conjugate $G = F^*$ of F is the function $\mathcal{X}'_F \subset \mathbb{R}^d \to \mathbb{R}$ defined by

$$F^*(x') = < x, x' > -F(x). \tag{20}$$



Primal (natural $\tilde{\Theta}$) Dual (expectation $\tilde{H}$)

Fig. 2. Clustering in the primal (natural) space $\tilde{\Theta}$ is dually equivalent to clustering in the dual (expectation) space $\tilde{H}$. The transformations are reversible. Both normal data sets are visualized in the source parameter space $\tilde{\Lambda}$.

It follows from Legendre transformation that *any* Bregman divergence D_F admits a *dual* Bregman divergence D_{F^*} related to D_F as follows:

$$D_F(p||q) = F(p) + F^*(\nabla F(q)) - < p, \nabla F(q) >, \tag{21}$$

$$= F(p) + F^*(q') - < p, q' >, \tag{22}$$

$$= D_{F^*}(q'||p'). \tag{23}$$

Yoshizawa and Tanabe [14] carried out non-trivial computations that yield the dual natural/expectation coordinate systems arising from the canonical decomposition of the density function $p(x; m, S)$:

$$\tilde{H} = \begin{pmatrix} \eta = \mu \\ H = -(\Sigma + \mu\mu^T) \end{pmatrix} \Longleftrightarrow \tilde{\Lambda} = \begin{pmatrix} \lambda = \mu \\ \Lambda = \Sigma \end{pmatrix}, \tag{24}$$

$$\tilde{\Lambda} = \begin{pmatrix} \lambda = \mu \\ \Lambda = \Sigma \end{pmatrix} \Longleftrightarrow \tilde{\Theta} = \begin{pmatrix} \theta = \Sigma^{-1}\mu \\ \Theta = \frac{1}{2}\Sigma^{-1} \end{pmatrix} \tag{25}$$

The strictly convex and differentiable dual Bregman generator functions (ie., potential functions in information geometry) are $F(\tilde{\Theta}) = \frac{1}{4}\mathrm{Tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log \det\Theta + \frac{d}{2}\log\pi$, and $F^*(\tilde{H}) = -\frac{1}{2}\log(1 + \eta^T H^{-1}\eta) - \frac{1}{2}\log\det(-H) - \frac{d}{2}\log(2\pi e)$ defined respectively both on the topologically open space $\mathbb{R}^d \times \mathcal{C}_d^-$, where $\mathcal{C}_d$ denote the d-dimensional cone of symmetric positive definite matrices. The $\tilde{H} \Leftrightarrow \tilde{\Theta}$ coordinate transformations obtained from the Legendre transformation are given by

$$\tilde{H} = \nabla_{\tilde{\Theta}} F(\tilde{\Theta}) = \begin{pmatrix} \nabla_{\tilde{\Theta}} F(\theta) \\ \nabla_{\tilde{\Theta}} F(\Theta) \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\Theta^{-1}\theta \\ -\frac{1}{2}\Theta^{-1} - \frac{1}{4}(\Theta^{-1}\theta)(\Theta^{-1}\theta)^T \end{pmatrix} \tag{26}$$

$$= \begin{pmatrix} \mu \\ -(\Sigma + \mu\mu^T) \end{pmatrix} \tag{27}$$

and

$$\tilde{\Theta} = \nabla_{\tilde{H}} F^*(\tilde{H}) = \begin{pmatrix} \nabla_{\tilde{H}} F^*(\eta) \\ \nabla_{\tilde{H}} F^*(H) \end{pmatrix} = \begin{pmatrix} -(H + \eta\eta^T)^{-1}\eta \\ -\frac{1}{2}(H + \eta\eta^T)^{-1} \end{pmatrix} = \begin{pmatrix} \Sigma^{-1}\mu \\ \frac{1}{2}\Sigma^{-1} \end{pmatrix}. \quad (28)$$

These formula simplify significantly when we restrict ourselves to diagonal-only variance-covariance matrices S_i, spherical Gaussians $S_i = s_i I$, or univariate normals $\mathcal{N}(m_i, s_i^2)$.

5 Left-Sided and Right-Sided Clusterings

The former Bregman k-means clustering makes use of the *right-side* of the divergence for clustering. It is therefore equivalent to the *left-side* clustering for the dual Bregman divergence on the gradient point set (see Figure 2). The left-side Kullback-Leibler clustering of members of the same exponential family is a right-side Bregman clustering for the log normalizer. Similarly, the right-side Kullback-Leibler clustering of members of the same exponential family is a *left-side* Bregman clustering for the log normalizer, that is itself equivalent to a right-side Bregman clustering for the dual convex conjugate $F*$ obtained from Legendre transformation.

We find that the left-side Bregman clustering (ie., right-side Kullback-Leibler) is *exactly* the clustering algorithm reported in [2]. In particular, the cluster centers for the right-side Kullback-Leibler divergence are left-side Bregman centroids that have been shown to be generalized means [15], given as (for $(\nabla\tilde{F})^{-1} = \nabla\tilde{F}^*$):

$$\tilde{\Theta} = (\nabla\tilde{F})^{-1}\left(\sum_{i=1}^{n} \nabla\tilde{F}(\tilde{\Theta}_i)\right). \quad (29)$$

After calculus, it follows in accordance with [2] that

$$S^* = \left(\frac{1}{n}\sum_i S_i^{-1}\right)^{-1}, \quad (30)$$

$$m^* = S^*(\sum_{i=1}^{n} \frac{1}{n} S_i^{-1} m_i). \quad (31)$$

6 Inferring Multivariate Normal Distributions

As mentioned in the introduction, in many real-world settings each datum point can be sampled several times yielding multiple observations assumed to be drawn from an underlying distribution. This modeling is convenient for considering individual noise characteristics. In many cases, we may also assume Gaussian sampling or Gaussian noise, see [2] for concrete examples in sensor data network and statistical debugging applications. The problem is then to infer from observations $x_1, ..., x_s$ the parameters m and S. It turns out that the maximum likelihood estimator (MLE) of exponential families is the centroid of the

sufficient statistics evaluated on the observations [7]. Since multivariate normal distributions belongs to the exponential families with statistics $(x, -\frac{1}{2}xx^T)$, it follows from the maximum likelihood estimator that

$$\hat{\mu} = \frac{1}{s}\sum_{i=1}^{s} x_i, \tag{32}$$

and

$$\hat{S} = \left(\frac{1}{2s}\sum_{i=1}^{s} x_i x_i^T\right) - \hat{\mu}\hat{\mu}^T. \tag{33}$$

This estimator may be *biased* [5].

7 Symmetric Clustering with the J-Divergence

The symmetrical Kullback-Leibler divergence $\frac{1}{2}(\mathrm{KL}(p||q)+\mathrm{KL}(q||p))$ is called the J-divergence. Although centroids for the left-side and right-side Kullback-Leibler divergence admit elegant closed-form solutions as *generalized means* [15], it is also known that the symmetrized Kullback-Leibler centroid of discrete distributions does not admit such a closed-form solution [16]. Nevertheless, the centroid of symmetrized Bregman divergence has been *exactly* geometrically characterized as the intersection of the geodesic linking the left- and right-sided centroids (say, c_L^F and c_R^F respectively) with the mixed-type bisector: $M_F(c_R^F, c_L^F) = \{x \in \mathcal{X} \mid D_F(c_R^F||x) = D_F(x||c_L^F)\}$. We summarize the geodesic-walk approximation heuristic of [15] as follows: We initially consider $\lambda \in [\lambda_m = 0, \lambda_M = 1]$ and repeat the following steps until $\lambda_M - \lambda_m \leq \epsilon$, for $\epsilon > 0$ a *prescribed* precision threshold:

1. **Geodesic walk.** Compute interval midpoint $\lambda_h = \frac{\lambda_m+\lambda_M}{2}$ and corresponding geodesic point

$$q_h = (\nabla F)^{-1}((1-\lambda_h)\nabla F(c_R^F) + \lambda_h \nabla F(c_L^F)), \tag{34}$$

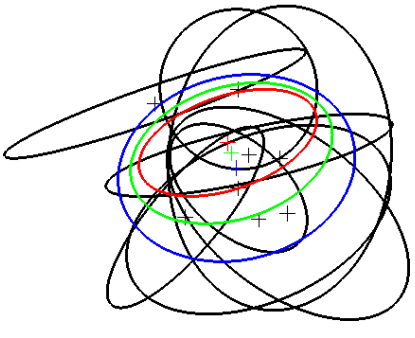


Fig. 3. The left-(red) and right-sided (blue) Kullback-Leibler centroids, and the symmetrized Kullback-Leibler J-divergence centroid (green) for a set of eight bivariate normals

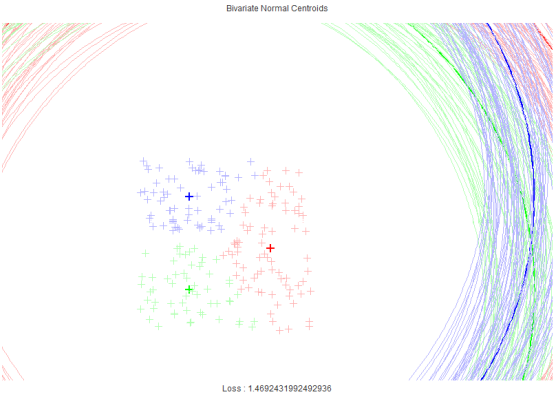


Fig. 4. Clustering sided or symmetrized multivariate normals. For identical variance-covariance matrices, this Bregman clustering amounts to the regular k-means. Indeed, in this case the Kullback-Leibler becomes proportional to the squared Euclidean distance. See demo applet at `http://www.sonycsl.co.jp/person/nielsen/KMj/`

2. **Mixed-type bisector side.** Evaluate the sign of $D_F(c_R^F||q_h) - D_F(q_h||c_L^R)$, and
3. **Dichotomy.** Branch on $[\lambda_h, \lambda_M]$ if the sign is negative, or on $[\lambda_m, \lambda_h]$ otherwise.

Figure 3 shows the two sided left- and right-sided centroids, and the symmetrized centroid for the case of bivariate normals (handled as points in 5D). We can then apply the classical k-means algorithm on these symmetrized centroids. Figure 4 displays that the multivariate clustering applet, which shows the property that it becomes the regular k-means if we fix all variance-covariance matrices to identity. See also the recent work of Teboulle [17] that further generalizes center-based clustering to Bregman and Csiszár f-divergences.

8 Concluding Remarks

We have presented the k-means hard clustering techniques [1] for clustering multivariate normals in arbitrary dimensions with respect to the Kullback-Leibler divergence. Our approach relies on instantiating the generic Bregman hard clustering of Banerjee et al. [9] by using the fact that the relative entropy between any two normal distributions can be derived from the corresponding mixed-type Bregman divergence obtained by setting the Bregman generator as the log normalizer function of the normal exponential family. This in turn yields a dual interpretation of the right-sided k-means clustering as a left-sided k-means clustering that was formerly studied by Davis and Dhillon [2] using an *ad-hoc* optimization technique. Furthermore, based on the very recent work on symmetrical Bregman centroids [15], we showed how to cluster multivariate normals with respect to the symmetrical Kullback-Leibler divergence, called the J-divergence.

This is all the more important for applications that require to handle symmetric information-theoretic measures [3].

References

1. Lloyd, S.P.: Least squares quantization in PCM. IEEE Transactions on Information Theory 28(2), 129–136 (1982); first published in 1957 in a Technical Note of Bell Laboratories
2. Davis, J.V., Dhillon, I.S.: Differential entropic clustering of multivariate gaussians. In: Scholkopf, B., Platt, J., Hoffman, T. (eds.) Neural Information Processing Systems (NIPS), pp. 337–344. MIT Press, Cambridge (2006)
3. Myrvoll, T.A., Soong, F.K.: On divergence-based clustering of normal distributions and its application to HMM adaptation. In: Proceedings of EuroSpeech, Geneva, Switzerland, vol. 2, pp. 1517–1520 (2003)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
5. Amari, S.I., Nagaoka, N.: Methods of Information Geometry. Oxford University Press, Oxford (2000)
6. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, Hoboken (2006)
7. Barndorff-Nielsen, O.E.: Parametric statistical models and likelihood. Lecture Notes in Statistics, vol. 50. Springer, New York (1988)
8. Nielsen, F., Boissonnat, J.D., Nock, R.: Bregman Voronoi diagrams: Properties, algorithms and applications, Extended abstract appeared in ACM-SIAM Symposium on Discrete Algorithms 2007. INRIA Technical Report RR-6154 (September 2007)
9. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. Journal of Machine Learning Research (JMLR) 6, 1705–1749 (2005)
10. Bregman, L.M.: The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics 7, 200–217 (1967)
11. Redmond, S.J., Heneghan, C.: A method for initialising the k-means clustering algorithm using kd-trees. Pattern Recognition Letters 28(8), 965–973 (2007)
12. Forgy, E.W.: Cluster analysis of multivariate data: efficiency vs interpretability of classifications. Biometrics 21, 768–769 (1965)
13. Carmi, P., Har-Peled, S., Katz, M.J.: On the Fermat-Weber center of a convex object. Computational Geometry 32(3), 188–195 (2005)
14. Yoshizawa, S., Tanabe, K.: Dual differential geometry associated with Kullback-Leibler information on the Gaussian distributions and its 2-parameter deformations. SUT Journal of Mathematics 35(1), 113–137 (1999)
15. Nielsen, F., Nock, R.: On the symmetrized Bregman centroids, Sony CSL Technical Report (submitted) (November 2007)
16. Veldhuis, R.N.J.: The centroid of the symmetrical Kullback-Leibler distance. IEEE Signal Processing Letters 9(3), 96–99 (2002)
17. Teboulle, M.: A unified continuous optimization framework for center-based clustering methods. Journal of Machine Learning Research 8, 65–102 (2007)