The Burbea-Rao and Bhattacharyya centroids

Frank Nielsen, *Senior Member, IEEE,* and Sylvain Boltz, *Nonmember, IEEE*

***Abstract*—We study the centroid with respect to the class of information-theoretic Burbea-Rao divergences that generalize the celebrated Jensen-Shannon divergence by measuring the non-negative Jensen difference induced by a strictly convex and differentiable function. Although those Burbea-Rao divergences are symmetric by construction, they are not metric since they fail to satisfy the triangle inequality. We first explain how a particular symmetrization of Bregman divergences called Jensen-Bregman distances yields exactly those Burbea-Rao divergences. We then proceed by defining skew Burbea-Rao divergences, and show that skew Burbea-Rao divergences amount in limit cases to compute Bregman divergences. We then prove that Burbea-Rao centroids can be arbitrarily finely approximated by a generic iterative concave-convex optimization algorithm with guaranteed convergence property. In the second part of the paper, we consider the Bhattacharyya distance that is commonly used to measure overlapping degree of probability distributions. We show that Bhattacharyya distances on members of the same statistical exponential family amount to calculate a Burbea-Rao divergence in disguise. Thus we get an efficient algorithm for computing the Bhattacharyya centroid of a set of parametric distributions belonging to the same exponential families, improving over former specialized methods found in the literature that were limited to univariate or "diagonal" multivariate Gaussians. To illustrate the performance of our Bhattacharyya/Burbea-Rao centroid algorithm, we present experimental performance results for k-means and hierarchical clustering methods of Gaussian mixture models.**

***Index Terms*—Centroid, Kullback-Leibler divergence, Jensen-Shannon divergence, Burbea-Rao divergence, Bregman divergences, Exponential families, Bhattacharrya divergence, Information geometry.**

I. Introduction

A. Means and centroids

In Euclidean geometry, the centroid c of a point set $\mathcal{P} = \{p_1, ..., p_n\}$ is defined as the center of mass $\frac{1}{n}\sum_{i=1}^n p_i$, also characterized as the center point that minimizes the *average squared* Euclidean distances: $c = \arg\min_p \sum_{i=1}^n \frac{1}{n}\|p - p_i\|^2$. This basic notion of Euclidean centroid can be extended to denote a *mean* point $M(\mathcal{P})$ representing the *centrality* of a given point set $\mathcal{P}$. There are basically two complementary approaches to define mean values of numbers: (1) by axiomatization, or (2) by optimization, summarized concisely as follows:

- **By axiomatization**. This approach was first historically pioneered by the independent work of Kolmogorov [1] and Nagumo [2] in 1930, and simplified and refined later by Aczél [3]. Without loss of generality we consider the mean of two non-negative numbers x_1 and x_2, and postulate the following expected behaviors of a mean function $M(x_1, x_2)$ as axioms (common sense):
  - Reflexivity. $M(x, x) = x$,
  - Symmetry. $M(x_1, x_2) = M(x_2, x_1)$,
  - Continuity and strict monotonicity. $M(\cdot, \cdot)$ continuous and $M(x_1, x_2) < M(x_1', x_2)$ for $x_1 < x_1'$, and
  - Anonymity. $M(M(x_{11}, x_{12}), M(x_{21}, x_{22})) = M(M(x_{11}, x_{21}), M(x_{12}, x_{22}))$ (also called bisymmetry expressing the fact that the mean can be computed as a mean on the row means or equivalently as a mean on the column means).

Then one can show that the mean function $M(\cdot, \cdot)$ is necessarily written as:

$$M(x_1, x_2) = f^{-1}\left(\frac{f(x_1) + f(x_2)}{2}\right) \stackrel{\mathrm{def}}{=} M_f(x_1, x_2), \quad (1)$$

for a strictly increasing function f. The arithmetic $\frac{x_1+x_2}{2}$, geometric $\sqrt{x_1 x_2}$ and harmonic means $\frac{2}{\frac{1}{x_1}+\frac{1}{x_2}}$ are instances of such generalized means obtained for $f(x) = x$, $f(x) = \log x$ and $f(x) = \frac{1}{x}$, respectively. Those generalized means are also called *quasi-arithmetic means*, since they can be interpreted as the arithmetic mean on the sequence $f(x_1), ..., f(x_n)$, the f-representation of numbers. To get geometric centroids, we simply consider means on each coordinate axis independently. The Euclidean centroid is thus interpreted as the Euclidean arithmetic mean. Barycenters (weighted centroids) are similarly obtained using non-negative weights (normalized so that $\sum_{i=1}^n w_i = 1$):

$$M_f(x_1, ..., x_n; w_1, ..., w_n) = f^{-1}\left(\sum_{i=1}^n w_i f(x_i)\right) \quad (2)$$

Those generalized means satisfy the inequality property:

$$M_f(x_1, ..., x_n; w_1, ..., w_n) \leq M_g(x_1, ..., x_n; w_1, ..., w_n), \quad (3)$$

if and only if function g dominates f: That is, $\forall x, g(x) > f(x)$. Therefore the arithmetic mean ($f(x) = x$) dominates the geometric mean ($f(x) = \log x$) which in turn dominates the harmonic mean $f(x) = \frac{1}{x}$. Note that it is *not* a strict inequality in Eq. 3 as the means coincide for all identical elements: if all x_i are equal to x then $M_f(x_1, ..., x_n) = f^{-1}(f(x)) = x = g^{-1}(g(x)) = M_g(x_1, ..., x_n)$. All those quasi-arithmetic means further satisfy the "interness" property

F. Nielsen is with the Department of Fundamental Research of Sony Computer Science Laboratories, Inc., Tokyo, Japan, and the Computer Science Department (LIX) of École Polytechnique, Palaiseau, France. e-mail: Frank.Nielsen@acm.org

S. Boltz is with the Computer Science Department (LIX) of École Polytechnique, Palaiseau, France. e-mail: boltz@lix.polytechnique.fr

Manuscript received April 2010

$$\min(x_1, ..., x_n) \leq M_f(x_1, ..., x_n) \leq \max(x_1, ..., x_n), \tag{4}$$

derived from limit cases $p \to \pm\infty$ of power means[1] for $f(x) = x^p, p \in \mathbb{R}_* = (-\infty, \infty)\backslash\{0\}$, a non-zero real number.

- **By optimization**. In this second alternative approach, the barycenter c is defined according to a distance function $d(\cdot,\cdot)$ as the optimal solution of a minimization problem

$$(\mathrm{OPT}) : \min_x \sum_{i=1}^n w_i d(x, p_i) = \min_x L(x; \mathcal{P}, d), \tag{5}$$

where the non-negative weights w_i denote multiplicity or relative importance of points (by default, the centroid is defined by fixing all $w_i = \frac{1}{n}$). Ben-Tal et al. [4] considered an information-theoretic class of distances called f-divergences [5], [6]:

$$I_f(x, p) = pf\left(\frac{x}{p}\right), \tag{6}$$

for a strictly convex differentiable function $f(\cdot)$ satisfying $f(1) = 0$ and $f'(1) = 0$. Although those f-divergences were primarily investigated for probability measures,[2] we can extend the f-divergence to positive measures. Since program (OPT) is *strictly convex* in x, it admits a *unique* minimizer $M(\mathcal{P}; I_f) = \arg\min_x L(x; \mathcal{P}, I_f)$, termed the *entropic mean* by Ben-Tal et al. [4]. Interestingly, those entropic means are linear scale-invariant:[3]

$$M(\lambda p_1, ..., \lambda p_n; I_f) = \lambda M(p_1, ..., p_n; I_f) \tag{7}$$

Nielsen and Nock [7] considered another class of information-theoretic distortion measures B_F called Bregman divergences [8], [9]:

$$B_F(x, p) = F(x) - F(p) - (x - p)F'(p), \tag{8}$$

for a strictly convex differentiable function F. It follows that (OPT) is convex, and admits a unique minimizer $M(p_1, ..., p_n; B_F) = M_{F'}(p_1, ..., p_n)$, a quasi-arithmetic mean for the strictly increasing and continuous function F', the derivative of F. Observe that information-theoretic distances may be asymmetric (i.e., $d(x, p) \neq d(p, x)$), and therefore one may also define a *right-sided* centroid M' as the minimizer of

$$(\mathrm{OPT}') : \min_x \sum_{i=1}^n w_i d(p_i, x), \tag{9}$$

It turns out that for f-divergences, we have:

$$I_f(x, p) = I_{f*}(p, x), \tag{10}$$

for $f^*(x) = xf(1/x)$ so that (OPT') is solved as a (OPT) problem for the *conjugate function* $f^*(\cdot)$. In the same spirit, we have:

$$B_F(x, p) = B_{F^*}(F'(p), F'(x)) \tag{11}$$

for Bregman divergences, where F^* denotes the Legendre convex conjugate [8], [9].[4] Surprisingly, although (OPT') may *not* be convex in x for Bregman divergences (e.g., $F(x) = -\log x$), (OPT') admits nevertheless a unique minimizer, independent of the generator function F: the center of mass $M'(\mathcal{P}; B_F) = \sum_{i=1}^n \frac{1}{n} p_i$. Bregman means are not homogeneous except for the power generators $F(x) = x^p$ which yields entropic means, i.e. means that can *also* be interpreted[5] as minimizers of average f-divergences [4]. Amari [11] further studied those power means (known as α-means in information geometry [12]), and showed that they are linear-scale free means obtained as minimizers of α-divergences, a proper subclass of f-divergences. Nielsen and Nock [13] reported an alternative simpler proof of α-means by showing that the α-divergences are Bregman divergences in disguise (namely, representational Bregman divergences for positive measures, but not for normalized distribution measures [10]). To get geometric centroids, we simply consider multivariate extensions of the optimization task (OPT). In particular, one may consider *separable* divergences that are divergences that can be assembled coordinate-wise:

$$d(x, p) = \sum_{i=1}^d d_i(x^{(i)}, p^{(i)}), \tag{12}$$

with $x^{(i)}$ denoting the ith coordinate. A typical non separable divergence is the squared Mahalanobis distance [14]:

$$d(x, p) = (x - p)^T Q(x - p), \tag{13}$$

a Bregman divergence called generalized quadratic distance, defined for the generator $F(x) = x^T Qx$, where Q is a positive-definite matrix ($Q \succ 0$). For separable distances, the optimization problem (OPT) may then be reinterpreted as the task of finding the projection [15] of a point p (of dimension $d \times n$) to the upper line U:

$$(\mathrm{PROJ}) : \inf_{u \in U} d(u, p) \tag{14}$$

with $u_1 = ... = u_{d \times n} > 0$, and p the $(n \times d)$-dimensional point obtained by stacking the d coordinates of each of the n points.

In geometry, means (centroids) play a crucial role in center-based clustering (i.e., k-means [16] for vector quantization applications). Indeed, the mean of a cluster allows one to *aggregate data* into a single center datum. Thus the notion

[1] Besides the min/max operators interpreted as extremal power means, the geometric mean itself can also be interpreted as a power mean $(\prod_{i=1}^n x_i^p)^{\frac{1}{p}}$ in the limit case $p \to 0$.

[2] In that context, a d-dimensional point is interpreted as a discrete and finite probability measure lying in the $(d-1)$-dimensional unit simplex.

[3] That is, means of homogeneous degree 1.

[4] Legendre dual convex conjugates F and F^* have necessarily reciprocal gradients: $F^{*\prime} = (F')^{-1}$. See [7].

[5] In fact, Amari [10] proved that the intersection of the class of f-divergences with the class of Bregman divergences are α-divergences.

of means are encapsulated into the broader theory of mathematical aggregators [17].

Results on geometric means can be easily transfered to the field of Statistics [4] by generalizing the optimization problem task to a random variable X with distribution F as:

$$(\mathrm{OPT}) : \min_x E[X d(x, X)] = \min_x \int_t t d(x, t) \mathrm{d}F(t), \quad (15)$$

where $E[\cdot]$ denotes the expectation defined with respect to the Lebesgue-Stieltjes integral. Although this approach is discussed in [4] and important for defining various notions of centrality in statistics, we shall not cover this extended framework here, for sake of brevity.

B. Burbea-Rao divergences

In this paper, we focus on the optimization approach (OPT) for defining other (geometric) means using the class of information-theoretic distances obtained by Jensen difference for a strictly convex and differentiable function F:

$$d(x, p) = \frac{F(x) + F(p)}{2} - F\left(\frac{x+p}{2}\right) \stackrel{\mathrm{def}}{=} \mathrm{BR}_F(x, p) \geq 0. \quad (16)$$

Since the underlying differential geometry implied by those Jensen difference distances have been seminally studied in papers of Burbea and Rao [18], [19], we shall term them Burbea-Rao divergences, and point out to them as BR_F. In the remainder, we consider separable Burbea-Rao divergences. That is, for d-dimensional points p and q, we define

$$\mathrm{BR}_F(p, q) = \sum_{i=1}^{d} \mathrm{BR}_F(p^{(i)}, q^{(i)}), \quad (17)$$

and study the Burbea-Rao centroids (and barycenters) as the minimizers of the average Burbea-Rao divergences. Those Burbea-Rao divergences generalize the celebrated Jensen-Shannon divergence [20]

$$\mathrm{JS}(p, q) = H\left(\frac{p+q}{2}\right) - \frac{H(p) + H(q)}{2} \quad (18)$$

by choosing $F(x) = -H(x)$, the negative Shannon entropy $H(x) = -x \log x$. Generators $F(\cdot)$ of parametric distances are convex functions representing entropies which are concave functions. Burbea-Rao divergences contain all generalized quadratic distances ($F(x) = x^T Q x = \langle Qx, x\rangle$ for a positive definite matrix $Q \succ 0$, also called squared Mahalanobis distances):

$$\begin{aligned} \mathrm{BR}_F(p, q) &= \frac{F(p) + F(q)}{2} - F\left(\frac{p+q}{2}\right) \\ &= \frac{2\langle Qp, p\rangle + 2\langle Qq, q\rangle - \langle Q(p+q), p+q\rangle}{4} \\ &= \frac{1}{4}(\langle Qp, p\rangle + \langle Qq, q\rangle - 2\langle Qp, q\rangle) \\ &= \frac{1}{4}\langle Q(p-q), p-q\rangle = \frac{1}{4}\|p-q\|_Q^2. \end{aligned}$$

Although the square root of the Jensen-Shannon divergence yields a metric (a Hilbertian metric), it is not true in general for Burbea-Rao divergences. The closest work to our paper is a 1-page symposium[6] paper [21] discussing about Ali-Silvey-Csiszár f-divergences [5], [6] and Bregman divergences [22], [8] (two entropy-based divergence classes). Those information-theoretic distortion classes are compared using quadratic differential metrics, mean values and projections. The notion of skew Jensen differences intervene in the discussion.

C. Contributions and paper organization

The paper is articulated into two parts: The first part studies the Burbea-Rao centroids, and the second part shows some applications in Statistics. We summarize our contributions as follows:

- We define the parametric class of (skew) Burbea-Rao divergences, and show that those divergences naturally arise when generalizing the principle of the Jensen-Shannon divergence [20] to Jensen-Bregman divergences. In the limit cases, we further prove that those skew Burbea-Rao divergences yield asymptotically Bregman divergences.
- We describe the centroids with respect to the (skew) Burbea-Rao divergences. Besides centroids for special cases of Burbea-Rao divergences (including the squared Euclidean distances), those centroids are not available in closed-form equations. However, we show that any Burbea-Rao centroid can be estimated efficiently using an iterative convex-concave optimization procedure. As a by-product, we find Bregman sided centroids [7] in closed-form in the extremal skew cases.

We then consider applications of Burbea-Rao centroids in Statistics, and show the link with Bhattacharyya distances. A wide class of statistical parametric models can be handled in a unified manner as exponential families [23]. The classes of exponential families contain many of the standard parametric models including the Poisson, Gaussian, multinomial, and Gamma/Beta distributions, just to name a few prominent members. However, only a few closed-form formulas for the statistical Bhattacharyya distances between those densities are reported in the literature.[7]

For the second part, our contributions are reviewed as follows:

- We show that the (skew) Bhattacharyya distances calculated for distributions belonging to the same exponential family in statistics, are equivalent to (skew) Burbea-Rao divergences. We mention corresponding closed-form formula for computing Chernoff coefficients and α-divergences of exponential families. In the limit case, we obtain an alternative proof showing that the Kullback-Leibler divergence of members of the same exponential

[6] In the nineties, the IEEE International Symposium on Information Theory (ISIT) published only 1-page papers. We are grateful to Prof. Michèle Basseville for sending us the corresponding slides.

[7] For instance, the Bhattacharyya distance between multivariate normal distributions is given here [24].

family is equivalent to a Bregman divergence calculated on the natural parameters [14].

- We approximate iteratively the Bhattacharyya centroid of any set of distributions of the same exponential family (including multivariate Gaussians) using the Burbea-Rao centroid algorithm. For the case of multivariate Gaussians, we design yet another tailored iterative scheme based on matrix differentials, generalizing the former univariate study of Rigazio et al. [25]. Thus we get either the generic way or the tailored way for computing the Bhattacharrya centroids of arbitrary Gaussians.
- As a field application, we show how to simplify Gaussian mixture models using hierarchical clustering, and show experimentally that the results obtained with the Bhattacharyya centroids compare favorably well with former results obtained for Bregman centroids [26]. Our numerical experiments show that the generic method outperforms the alternative tailored method for multivariate Gaussians.

The paper is organized as follows: In section II, we introduce Burbea-Rao divergences as a natural extension of the Jensen-Shannon divergence using the framework of Bregman divergences. It is followed by Section III which considers the general case of skew divergences, and reveals asymptotic behaviors of extreme skew Burbea-Rao divergences as Bregman divergences. Section IV defines the (skew) Burbea-Rao centroids, and present a simple iterative algorithm with guaranteed convergence. We then consider applications in Statistics in Section V: After briefly recalling exponential distributions in §V-A, we show that Bhattacharyya distances and Chernoff/Amari α-divergences are available in closed-form equations as Burbea-Rao divergences for distributions of the same exponential families. Section V-C presents an alternative iterative algorithm tailored to compute the Bhattacharyya centroid of multivariate Gaussians, generalizing the former specialized work of Rigazio et al. [25]. In section V-D, we use those Bhattacharyya/Burbea-Rao centroids to simplify hierarchically Gaussian mixture models, and comment both qualitatively and quantitatively our experiments on a color image segmentation application. Finally, section VI concludes this paper by describing further perspectives and hinting at some information geometrical aspects of this work.

II. Burbea-Rao divergences from symmetrization of Bregman divergences

Let $\mathbb{R}^+ = [0, +\infty)$ denote the set of non-negative reals. For a strictly convex (and differentiable) generator F, we define the Burbea-Rao divergence as the following non-negative function:

$$\begin{aligned} \mathrm{BR}_F \quad &: \quad \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+ \\ (p,q) \quad &\mapsto \quad \mathrm{BR}_F(p,q) = \frac{F(p)+F(q)}{2} - F\left(\frac{p+q}{2}\right) \geq 0 \end{aligned}$$

The non-negative property of those divergences follows straightforwardly from Jensen inequality. Although Burbea-Rao distances are symmetric ($\mathrm{BR}_F(p,q) = \mathrm{BR}_F(q,p)$), they are not metrics since they fail to satisfy the triangle inequality. A geometric interpretation of those divergences is given in Figure 1. Note that F is defined up to an affine term $ax+b$.

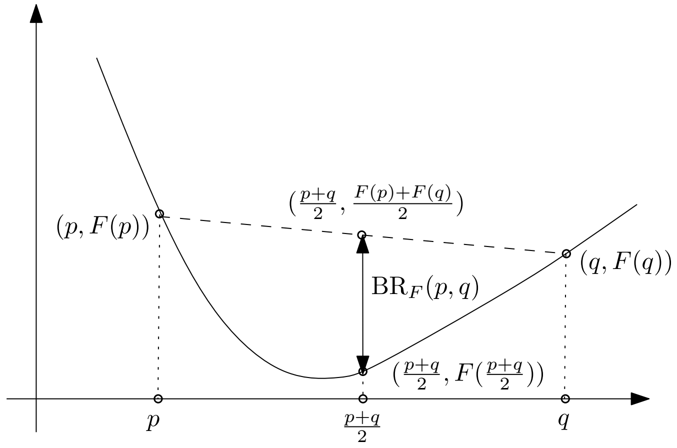
$(p, F(p))$
$(\frac{p+q}{2}, \frac{F(p)+F(q)}{2})$
$(q, F(q))$
$\mathrm{BR}_F(p,q)$
$(\frac{p+q}{2}, F(\frac{p+q}{2}))$
p
$\frac{p+q}{2}$
q

Fig. 1. Interpreting the Burbea-Rao divergence $\mathrm{BR}_F(p,q)$ as the vertical distance between the midpoint of segment $[(p,F(p)),(q,F(q))]$ and the midpoint of the graph plot $\left(\frac{p+q}{2}, F\left(\frac{p+q}{2}\right)\right)$.

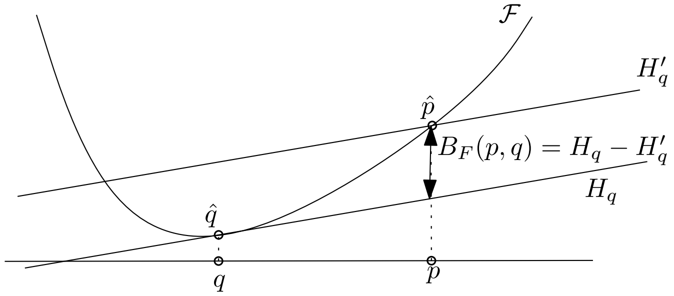

Fig. 2. Interpreting the Bregman divergence $B_F(p,q)$ as the vertical distance between the tangent plane at q and its translate passing through p (with identical slope $\nabla F(q)$).

We show that Burbea-Rao divergences extend the Jensen-Shannon divergence using the broader concept of Bregman divergences instead of the Kullback-Leibler divergence. A Bregman divergence [22], [8], [9] B_F is defined as the positive tail of the first-order Taylor expansion of a strictly convex and differentiable convex function F:

$$B_F(p,q) = F(p) - F(q) - \langle p-q, \nabla F(q)\rangle, \qquad (19)$$

where ∇F denote the gradient of F (the vector of partial derivatives $\{\frac{\partial F}{\partial x_i}\}_i$), and $\langle x,y\rangle = x^T y$ the inner product (dot product for vectors). A Bregman divergence is interpreted geometrically [14] as the vertical distance between the tangent plane H_q at q of the graph plot $\mathcal{F} = \{\hat{x} = (x, F(x)) \mid x \in \mathcal{X}\}$ and its translates H'_q passing through $\hat{p} = (p, F(p))$. Figure 2 depicts graphically the geometric interpretation of the Bregman divergence (to be compared with the Burbea-Rao divergence in Figure 1).

Bregman divergences are never metrics, and symmetric only for the generalized quadratic distances [14] obtained by choosing $F(x) = x^T Q x$, for some positive definite matrix $Q \succ 0$. Bregman divergences allow one to encapsulate both statistical distances with geometric distances:

- Kullback-Leibler divergence obtained for $F(x) = x\log x$:
$$\mathrm{KL}(p,q) = \sum_{i=1}^{d} p^{(i)} \log \frac{p^{(i)}}{q^{(i)}} \tag{20}$$
- squared Euclidean distance obtained for $F(x) = x^2$:
$$L_2^2(p,q) = \sum_{i=1}^{d} (p^{(i)} - q^{(i)})^2 = \|p-q\|^2 \tag{21}$$

Basically, there are two ways to symmetrize Bregman divergences (see also work on Bregman metrization [27], [28]):

- **Jeffreys-Bregman divergences.** We consider half of the double-sided divergences:
$$\begin{aligned} S_F(p;q) &= \frac{B_F(p,q)+B_F(q,p)}{2} & (22)\\ &= \frac{1}{2}\langle p-q, \nabla F(p) - \nabla F(q)\rangle, & (23)\end{aligned}$$
Except for the generalized quadratic distances, this symmetric distance *cannot* be interpreted as a Bregman divergence [14].
- **Jensen-Bregman divergences.** We consider the Jeffreys-Bregman divergences from the source parameters to the average parameter $\frac{p+q}{2}$ as follows:
$$\begin{aligned} J_F(p;q) &= \frac{B_F(p,\frac{p+q}{2}) + B_F(q,\frac{p+q}{2})}{2} & (24)\\ &= \frac{F(p)+F(q)}{2} - F(\frac{p+q}{2}) = \mathrm{BR}_F(p,q) \end{aligned}$$

Note that even for the negative Shannon entropy $F(x) = x\log x - x$ (extended to positive measures), those two symmetrizations yield different divergences: While S_F uses the gradient ∇F, J_F relies only on the generator F. Both J_F and S_F have always finite values.[8] The first symmetrization approach was historically studied by Jeffreys [29].

The second way to symmetrize Bregman divergences generalizes the spirit of the Jensen-Shannon divergence [20]

$$\begin{aligned} \mathrm{JS}(p,q) &= \frac{1}{2}\left(\mathrm{KL}\left(p,\frac{p+q}{2}\right) + \mathrm{KL}\left(q,\frac{p+q}{2}\right)\right)\!, & (25)\\ &= H\left(\frac{p+q}{2}\right) - \frac{H(p)+H(q)}{2} & (26)\end{aligned}$$

with non-negativity that can be derived from Jensen's inequality, hence its name. The Jensen-Shannon divergence is also called the total divergence to the average, a generalized measure of diversity from the *population distributions* p and q to the *average population* $\frac{p+q}{2}$. Those Jensen difference-type divergences are by definition Burbea-Rao divergences. For the Shannon entropy, those two different information divergence symmetrizations (Jensen-Shannon divergence and Jeffreys J divergence) satisfy the following inequality:

$$J(p,q) \geq 4\ \mathrm{JS}(p,q) \geq 0. \tag{27}$$

Nielsen and Nock [7] investigated the centroids with respect to Jeffreys-Bregman divergences (the symmetrized Kullback-Leibler divergence).

III. Skew Burbea-Rao divergences

We further generalize Burbea-Rao divergences by introducing a positive weight $\alpha \in (0,1)$ when averaging source parameters p and q as follows:

$$\begin{aligned} \mathrm{BR}_F^{(\alpha)} &: \mathcal{X}\times\mathcal{X} \to \mathbb{R}^+ \\ \mathrm{BR}_F^{(\alpha)}(p,q) &= \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q)\end{aligned}$$

We consider the open interval $(0,1)$ since otherwise the divergence has no discriminatory power (indeed, for $\alpha \in \{0,1\}, \mathrm{BR}_F^{(\alpha)}(p,q) = 0,\ \forall p,q$). Although skewed divergences are asymmetric $\mathrm{BR}_F^{(\alpha)}(p,q) \neq \mathrm{BR}_F^{(\alpha)}(q,p)$, we can swap arguments by replacing α by $1-\alpha$:

$$\begin{aligned} \mathrm{BR}_F^{(\alpha)}(p,q) &= \alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q) \\ &= \mathrm{BR}_F^{(1-\alpha)}(q,p) & (28)\end{aligned}$$

Those skew Burbea-Rao divergences are similarly found using a skew Jensen-Bregman counterpart (the gradient terms $\nabla F(\alpha p + (1-\alpha)q)$ perfectly cancel in the sum of skew Bregman divergences):

$$\alpha B_F(p, \alpha p + (1-\alpha)q) + (1-\alpha)B_F(q, \alpha p + (1-\alpha)q) \stackrel{\mathrm{def}}{=} \mathrm{BR}_F^{(\alpha)}(p,q)$$

In the limit cases, $\alpha \to 0$ or $\alpha \to 1$, we have $\mathrm{BR}_F^{(\alpha)}(p,q) \to 0\ \forall p,q$. That is, those divergences loose their discriminatory power at extremities. However, we show that those skew Burbea-Rao divergences tend *asymptotically* to Bregman divergences:

$$\begin{aligned} B_F(p,q) &= \lim_{\alpha\to 0} \frac{1}{\alpha}\mathrm{BR}_F^{(\alpha)}(p,q) & (29)\\ B_F(q,p) &= \lim_{\alpha\to 1} \frac{1}{1-\alpha}\mathrm{BR}_F^{(\alpha)}(p,q) & (30)\end{aligned}$$

The limit in the right-hand-side of Eq. 30 can be expressed alternatively as the following one-sided limit:

$$\lim_{\alpha\uparrow 1} \frac{1}{1-\alpha}\mathrm{BR}_F^{(\alpha)}(p,q) = \lim_{\alpha\downarrow 0} \frac{1}{\alpha}\mathrm{BR}_F^{(\alpha)}(q,p), \tag{31}$$

where the arrows ↑ and ↓ denote the limit from the left and the limit from the right, respectively (see [30] for notations). The right derivative of a function f at x is defined as $f'_+(x) = \lim_{y\downarrow x} \frac{f(y)-f(x)}{y-x}$. Since $\mathrm{BR}_F^{(0)}(p,q) = 0\ \forall p,q$, it follows that the right-hand-side limit of Eq. 31 is the right derivative (see Theorem 1 of [30] that gives a generalized Taylor expansion of convex functions) of the map

$$L(\alpha) : \alpha \mapsto \mathrm{BR}_F^{(\alpha)}(q,p) \tag{32}$$

taken at $\alpha = 0$. Thus we have

[8] This may not be the case of Bregman/Kullback-Leibler divergences that can potentially be unbounded.

$$\lim_{\alpha\downarrow 0}\frac{1}{\alpha}\mathrm{BR}_F^{(\alpha)}(q,p) = L'_+(0).,\qquad (33)$$

with

$$\begin{aligned} L'_+(0) &= \frac{\mathrm{d}_+}{\mathrm{d}\alpha}(\alpha F(q) + (1-\alpha)F(p) - F(\alpha q + (1-\alpha)p)) \\ &= F(q) - F(p) - \langle q-p, \nabla F(p)\rangle \qquad (34)\\ &= B_F(q,p) \qquad (35)\end{aligned}$$

Lemma 1: Skew Burbea-Rao divergences tend asymptotically to Bregman divergences ($\alpha \rightarrow 0$) or reverse Bregman divergences ($\alpha \rightarrow 1$).

Thus we may scale skew Burbea-Rao divergences so that Bregman divergences belong to skew Burbea-Rao divergences:

$$\begin{aligned} &\mathrm{sBR}_F^{(\alpha)}(p,q) = \\ &\quad \frac{1}{\alpha(1-\alpha)}\left(\alpha F(p) + (1-\alpha)F(q) - F(\alpha p + (1-\alpha)q)\right) \end{aligned} \qquad (36)$$

Moreover, α is now not anymore restricted to $(0,1)$ but to the full real line: $\alpha \in \mathbb{R}$, as also noticed in [31]. Setting $\alpha = \frac{1-\alpha'}{2}$ (that is, $\alpha' = 1-2\alpha$), we get

$$\begin{aligned} &\mathrm{sBR}_F^{(\alpha')}(p,q) = \\ &\quad \frac{4}{1-\alpha'^2}\left(\frac{1-\alpha'}{2}F(p) + \frac{1+\alpha'}{2}F(q) - F\left(\frac{1-\alpha'}{2}p + \frac{1+\alpha'}{2}q\right)\right) \end{aligned} \qquad (37)$$

IV. Burbea-Rao centroids

Let $\mathcal{P} = \{p_1, ..., p_n\}$ denote a d-dimensional point set. To each point, let us further associate a positive weight w_i (accounting for arbitrary multiplicity) and a positive scalar $\alpha_i \in (0,1)$ to define an anchored distance $\mathrm{BR}_F^{(\alpha_i)}(\cdot, p_i)$. Define the skew Burbea-Rao barycenter (or centroid) c as the minimizer of the following optimization task:

$$\mathrm{OPT} : c = \arg\min_x \sum_{i=1}^n w_i \mathrm{BR}_F^{(\alpha_i)}(x, p_i) = \arg\min_x L(x) \qquad (38)$$

Without loss of generality, we consider argument x on the left argument position (otherwise, we change all $\alpha_i \rightarrow 1-\alpha_i$ to get the right-sided Burbea-Rao centroid). Removing all terms independent of x, the minimization program (OPT) amounts to minimize equivalently the following energy function:

$$E(c) = (\sum_{i=1}^n w_i\alpha_i)F(c) - \sum_{i=1}^n w_i F(\alpha_i c + (1-\alpha_i)p_i) \qquad (39)$$

Observe that the energy function is decomposable in the sum of a convex function $(\sum_{i=1}^n w_i\alpha_i)F(c)$ with a concave function $-\sum_{i=1}^n w_i F(\alpha_i c + (1-\alpha_i)p_i)$ (since the sum of n concave functions is concave). We can thus solve iteratively this optimization problem using the Convex-ConCave Procedure [32], [33] (CCCP), by starting from an initial position c_0 (say, the barycenter $c_0 = \sum_{i=1}^n w_i p_i$), and iteratively update the barycenter as follows:

$$\nabla F(c_{t+1}) = \frac{1}{\sum_{i=1}^n w_i\alpha_i}\sum_{i=1}^n w_i\alpha_i \nabla F\left(\alpha_i c_t + (1-\alpha_i)p_i\right) \qquad (40)$$

$$c_{t+1} = \nabla F^{-1}\left(\frac{1}{\sum_{i=1}^n w_i\alpha_i}\sum_{i=1}^n w_i\alpha_i \nabla F\left(\alpha_i c_t + (1-\alpha_i)p_i\right)\right) \qquad (41)$$

Since F is convex, the second-order derivative $\nabla^2 F$ is always positive definite, and ∇F is strictly monotone increasing. Thus we can interpret Eq. 41 as a fixed-point equation by considering the ∇F-representation. Each iteration is interpreted as a quasi-arithmetic mean. This proves that the Burbea-Rao centroid is always well-defined. There is (at most) a unique fixed point for $x = g(x)$ with a function $g(\cdot)$ strictly monotone increasing.

In some cases, like the squared Euclidean distance (or squared Mahalanobis distances), we find closed-form solutions for the Burbea-Rao barycenters. For example, consider the (negative) quadratic entropy $F(x) = \langle x, x\rangle = \sum_{i=1}^d (x^{(i)})^2$ with weights w_i and all $\alpha_i = \frac{1}{2}$ (non-skew symmetric Burbea-Rao divergences). We have:

$$\begin{aligned} \min E(x) &= \frac{F(x)}{2} - \sum_{i=1}^n w_i F\left(\frac{p_i + x}{2}\right), \qquad (42)\\ &= \min \frac{\langle x,x\rangle}{2} - \frac{1}{4}\sum_{i=1}^n w_i\left(\langle x,x\rangle + 2\langle x,p_i\rangle + \langle p_i,p_i\rangle\right) \end{aligned}$$

The minimum is obtained when the gradient $\nabla E(x) = 0$, that is when $x = \bar{p} = \sum_{i=1}^n w_i p_i$, the barycenter of the point set $\mathcal{P}$. For most Burbea-Rao divergences, Eq. 42 can only be solved numerically.

Observe that for extremal skew cases (for $\alpha \rightarrow 0$ or $\alpha \rightarrow 1$), we obtain the Bregman centroids in closed-form solutions (see Eq. 30). Thus skew Burbea-Rao centroids allow one to get a smooth transition from the right-sided centroid (the center of mass) to the left-sided centroid (a quasi-arithmetic mean M_f obtained for $f = \nabla F$, a continuous and strictly increasing function).

Theorem 1: Skew Burbea-Rao centroids can be estimated iteratively using the CCCP iterative algorithm. In extremal skew cases, the Burbea-Rao centroids tend to Bregman left/right sided centroids, and have closed-form equations in limit cases.

To describe the orbit of Burbea-Rao centroids linking the left to right sided Bregman centroids, we compute for $\alpha \in [0,1]$ the skew Burbea-Rao centroids with the following update scheme:

$$c_{t+1} = \nabla F^{-1}\left(\sum_{i=1}^n w_i \nabla F(\alpha c_t + (1-\alpha)p_i)\right) \qquad (43)$$

We may further consider various convex generators F_i for each point, and consider the updating scheme

$$c_{t+1} = \left(\sum_i w_i \nabla F_i\right)^{-1} \left(\frac{1}{\sum_{i=1}^n w_i \alpha_i} \sum_{i=1}^n w_i \alpha_i \nabla F_i(\alpha_i c_t + (1-\alpha_i) p_i)\right)$$

A. Burbea-Rao divergences of a population

Consider now the Burbea-Rao divergence of a population $p_1, ..., p_n$ with respective positive normalized weights $w_1, ..., w_n$. The Burbea-Rao divergence is defined by:

$$\mathrm{BR}_F^w(p_1, ..., p_n) = \sum_{i=1}^n w_i F(p_i) - F(\sum_{i=1}^n w_i p_i) \geq 0 \qquad (44)$$

This family of diversity measures includes the Jensen-Rényi divergences [34], [35] for $F(x) = -R_\alpha(x)$, where $R_\alpha(x) = \frac{1}{1-\alpha} \log \sum_{j=1}^d p_j^\alpha$ is the Rényi entropy of order α. (Rényi entropy is concave for $\alpha \in (0,1)$ and tend to Shannon entropy for $\alpha \rightarrow 1$.)

V. Bhattacharyya distances as Burbea-Rao distances

We first briefly recall the versatile class of exponential family distributions in Section V-A. Then we show in Section V-B that the statistical Bhattacharyya/Chernoff distances between exponential family distributions amount to compute a Burbea-Rao divergence.

A. Exponential family distribution in Statistics

Many usual statistical parametric distributions $p(x;\lambda)$ (e.g., Gaussian, Poisson, Bernoulli/multinomial, Gamma/Beta, etc.) share common properties arising from their common canonical decomposition of probability distribution [9]:

$$p(x;\lambda) = p_F(x;\theta) = \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right). \qquad (45)$$

Those distributions[9] are said to belong to the exponential families (see [23] for a tutorial). An exponential family is characterized by its *log-normalizer* $F(\theta)$, and a distribution in that family by its *natural parameter* θ belonging to the *natural space* Θ. The log-normalizer F is strictly convex and C^∞, and can also be expressed using the source coordinate system λ using the 1-to-1 map $\tau : \Lambda \rightarrow \Theta$ that converts parameters from the source coordinate system λ to the natural coordinate system θ:

$$F(\theta) = F(\tau(\lambda)) = (F \circ \tau)(\lambda) = F_\lambda(\lambda), \qquad (46)$$

where $F_\lambda = F \circ \tau$ denotes the log-normalizer function expressed using the λ-coordinates instead of the natural θ-coordinates.

The vector $t(x)$ denote the *sufficient statistics*, that is the set of linear independent functions that allows to concentrate without any loss all information about the parameter θ carried in the iid. observations $x_1, x_2, ...,$. The inner product $\langle p, q\rangle$ is defined according to the primitive type of θ. Namely, it is a multiplication $\langle p, q\rangle = pq$ for scalars, a dot product $\langle p, q\rangle = p^T q$ for vectors, a matrix trace $\langle p, q\rangle = \mathrm{tr}(p^T \times q) = \mathrm{tr}(p \times q^T)$ for matrices, etc. For composite types such as p being defined by both a vector part and a matrix part, the composite inner product is defined as the sum of inner products on the primitive types. Finally, $k(x)$ represents the carrier measure according to the counting or Lebesgue measures. Decompositions for most common exponential family distributions are given in [23]. An exponential family $\mathcal{E}_F = \{p_F(x;\theta)\ |\theta \in \Theta\}$ is the set of probability distributions obtained for the same log-normalizer function F. Information geometry considers $\mathcal{E}_F$ as a manifold entity, and study its differential geometric properties [12].

For example, consider the family of Poisson distributions $\mathcal{E}_F$ with mass function:

$$p(x;\lambda) = \frac{\lambda^x}{x!} \exp(-\lambda), \qquad (47)$$

for $x \in \mathbb{N}_+ = \mathbb{N} \cup \{0\}$ a positive integer. Poisson distributions are univariate exponential families ($x \in \mathbb{N}_+$) of order 1 (parameter λ). The canonical decomposition yields

- the sufficient statistic $t(x) = x$,
- $\theta = \log \lambda$, the natural parameter,
- $F(\theta) = \exp \theta$, the log-normalizer,
- and $k(x) = -\log x!$ the carrier measure (with respect to the counting measure).

Since we deal with applications using multivariate normals in the following, we also report explicitly that canonical decomposition for the multivariate Gaussian family $\{p_F(x;\theta)\ |\theta \in \Theta\}$. We rewrite the usual Gaussian density of mean μ and variance-covariance matrix Σ:

$$\begin{aligned} p(x;\lambda) &= p(x;\mu,\Sigma) && (48)\\ &= \frac{1}{2\pi\sqrt{\det\Sigma}} \exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu))}{2}\right) && (49)\end{aligned}$$

in the canonical form of Eq. 45 with,

- $\theta = (\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}) \in \Theta = \mathbb{R}^d \times \mathbb{K}_{d\times d}$, with $\mathbb{K}_{d\times d}$ denotes the cone of positive definite matrices,
- $F(\theta) = \frac{1}{4}\mathrm{tr}(\theta_2^{-1}\theta_1\theta_1^T) - \frac{1}{2}\log\det\theta_2 + \frac{d}{2}\log\pi$,
- $t(x) = (x, -x^T x)$,
- $k(x) = 0$.

In this case, the inner product is composite and is calculated as the sum of a dot product and a matrix trace as follows:

$$\langle\theta, \theta'\rangle = \theta_1^T\theta_1' + \mathrm{tr}(\theta_2^T\theta_2'). \qquad (50)$$

The coordinate transformation $\tau : \Lambda \rightarrow \Theta$ is given for $\lambda = (\mu, \Sigma)$ by

$$\tau(\lambda) = \left(\lambda_2^{-1}\lambda_1, \frac{1}{2}\lambda_2^{-1}\right), \qquad (51)$$

and its inverse mapping $\tau^{-1} : \Theta \rightarrow \Lambda$ by

$$\tau^{-1}(\theta) = \left(\frac{1}{2}\theta_2^{-1}\theta_1, \frac{1}{2}\theta_2^{-1}\right). \qquad (52)$$

[9] The distributions can either be discrete or continuous. We do not introduce the unifying framework of probability measures in order to not burden the paper.

B. Bhattacharyya/Chernoff coefficients and α-divergences as skew Burbea-Rao divergences

For arbitrary probability distributions $p(x)$ and $q(x)$ (parametric or not), we measure the amount of overlap between those distributions using the Bhattacharyya coefficient [36]:

$$C_(p,q) = \int \sqrt{p(x)q(x)}\mathrm{d}x, \tag{53}$$

Clearly, the Bhattacharyya coefficient (measuring the affinity between distributions [37]) falls in the unit range:

$$0 \leq C(p,q) \leq 1. \tag{54}$$

In fact, we may interpret this coefficient geometrically by considering $\sqrt{p(x)}$ and $\sqrt{q(x)}$ as unit vectors. The Bhattacharyya distance is then the dot product, representing the cosine of the angle made by the two unit vectors. The Bhattacharyya distance $B : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is derived from its coefficient [36] as

$$B(p,q) = -\ln C(p,q). \tag{55}$$

The Bhattacharyya distance allows one to get *both* upper and lower bound the Bayes' classification error [38], [39], while there are no such results for the symmetric Kullback-Leibler divergence. Both the Bhattacharyya distance and the symmetric Kullback-Leibler divergence agrees with the Fisher information at the infinitesimal level. Although the Bhattacharyya distance is symmetric, it is not a metric. Nevertheless, it can be metrized by transforming it into to the following Hellinger metric [40]:

$$H(p,q) = \sqrt{\frac{1}{2}\int(\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}x}, \tag{56}$$

such that $0 \leq H(p,q) \leq 1$. It follows that

$$\begin{aligned} H(p,q) &= \\ & \sqrt{\frac{1}{2}\left(\int p(x)\mathrm{d}x + \int q(x)\mathrm{d}x - 2\int \sqrt{p(x)}\sqrt{q(x)}\mathrm{d}x\right)} \\ = & \sqrt{1 - C(p,q)}. \end{aligned} \tag{57}$$

Hellinger metric is also called Matusita metric [37] in the literature. The thesis of Hellinger was emphasized in the work of Kakutani [41].

We consider a direct generalization of Bhattacharyya coefficients and divergences called Chernoff divergences[10]

$$\begin{aligned} B_\alpha(p,q) &= -\ln \int_x p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x = -\ln C_\alpha(p,q) & (58) \\ &= -\ln \int_x q(x)\left(\frac{p(x)}{q(x)}\right)^\alpha \mathrm{d}x & (59) \\ &= -\ln E_q[L^\alpha(x)] & (60) \end{aligned}$$

defined for some $\alpha \in (0,1)$ (the Bhattacharyya divergence is obtained for $\alpha = \frac{1}{2}$), where $E[\cdot]$ denote the expectation, and $L(x) = \frac{p(x)}{q(x)}$ the likelihood ratio. The term $\int_x p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x$ is called the Chernoff coefficient. The Bhattacharyya/Chernoff distance of members of the same exponential family yields a weighted asymmetric Burbea-Rao divergence (namely, a *skew* Burbea-Rao divergence):

$$B_\alpha(p_F(x;\theta_p), p_F(x;\theta_q)) = \mathrm{BR}_F^{(\alpha)}(\theta_p, \theta_q) \tag{61}$$

with

$$\mathrm{BR}_F^{(\alpha)}(\theta_p,\theta_q) = \alpha F(\theta_p) + (1-\alpha)F(\theta_q) - F(\alpha\theta_p + (1-\alpha)\theta_q) \tag{62}$$

Chernoff coefficients are also related to α-divergences, the canonical divergences in α-flat spaces in information geometry [12] (p. 57):

$$D_\alpha(p||q) = \begin{cases} \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}}\mathrm{d}x\right), & \alpha \neq \pm 1, \\ \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x = \mathrm{KL}(p,q), & \alpha = -1, \\ \int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}x = \mathrm{KL}(q,p), & \alpha = 1, \end{cases} \tag{63}$$

The class of α-divergences satisfy the following reference duality: $D_\alpha(p||q) = D_{-\alpha}(q||p)$. Remapping $\alpha' = \frac{1-\alpha}{2}$ ($\alpha = 1 - 2\alpha'$), we transform Amari α-divergences to Chernoff α'-divergences:[11]

$$D_{\alpha'}(p,q) = \begin{cases} \frac{1}{\alpha'(1-\alpha')}\left(1 - \int p(x)^{\alpha'} q(x)^{1-\alpha'}\mathrm{d}x\right), & \alpha' \not\in \{0,1\}, \\ \int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x = \mathrm{KL}(p,q), & \alpha' = 1, \\ \int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}x = \mathrm{KL}(q,p), & \alpha' = 0, \end{cases} \tag{64}$$

Theorem 2: The Chernoff α'-divergence ($\alpha \neq \pm 1$) of distributions belonging to the same exponential family is given in closed-form by means of a skewed Burbea-Rao divergence as: $D_{\alpha'}(p,q) = \frac{1}{\alpha'(1-\alpha')}(1 - e^{-\mathrm{BR}_F^{\alpha'}(\theta_p,\theta_q)})$, with $\mathrm{BR}_F^{(\alpha)}(\theta_p,\theta_q) = (\alpha F(\theta_p) - (1-\alpha)F(\theta_q)) - F(\alpha\theta_p - (1-\alpha)\theta_q)$. Amari α-divergence for members of the same exponential families amount to compute $D_\alpha(p,q) = \frac{4}{1-\alpha^2}(1 - e^{-\mathrm{BR}_F^{\left(\frac{1-\alpha}{2}\right)}(\theta_p,\theta_q)})$

We get the following theorem for Bhattacharyya/Chernoff distances:

Theorem 3: The skew Bhattacharyya divergence $B_\alpha(p,q)$ is equivalent to the Burbea-Rao divergence for members of the same exponential family $\mathcal{E}_F$: $B_\alpha(p,q) = B_\alpha(p_F(x;\theta_p), p_F(x;\theta_q)) = -\log C_\alpha(p_F(x;\theta_p), p_F(x;\theta_q)) = \mathrm{BR}_F^{(\alpha)}(\theta_p,\theta_q) \geq 0$.

In particular, for $\alpha = \pm 1$, the Kullback-Leibler divergence of those exponential family distributions amount to compute

[10] In the literature, Chernoff information is also defined as $-\log\inf_{\alpha\in[0,1]}\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x$. Similarly, Chernoff coefficients $C_\alpha(p,q)$ are defined as the supremum: $C_\alpha(p,q) = \sup_{\alpha\in[0,1]}\int p^\alpha(x)q^{1-\alpha}(x)\mathrm{d}x$.

[11] Chernoff coefficients are also related to Rényi α-divergence generalizing the Kullback-Leibler divergence: $R_\alpha(p||q) = \frac{1}{\alpha-1}\log\int_x p(x)^\alpha q^{1-\alpha}(x)\mathrm{d}x$ built on Rényi entropy $H_R^\alpha(p) = \frac{1}{1-\alpha}\log(\int_x p^\alpha(x)\mathrm{d}x - 1)$. The Tsallis entropy $H_T^\alpha(p) = \frac{1}{\alpha-1}(1 - \int p(x)^\alpha \mathrm{d}x)$ can also be obtained from the Rényi entropy (and vice-versa) via the mappings: $H_T^\alpha(p) = \frac{1}{1-\alpha}(e^{(1-\alpha)H_R^\alpha(p)} - 1)$ and $H_R^\alpha(p) = \frac{1}{1-\alpha}\log(1 + (1-\alpha)H_T^\alpha(p))$.

Let us compute the Chernoff coefficient for distributions belonging to the same exponential families. Without loss of generality, let us consider the reduced canonical form of exponential families $p_F(x;\theta)=\exp\langle x,\theta\rangle - F(\theta)$. Chernoff coefficients $C_\alpha(p,q)$ of members $p=p_F(x;\theta_p)$ and $q=p_F(x;\theta_q)$ of the *same* exponential family $\mathcal{E}_F$:

$$\begin{aligned}
C_\alpha(p,q) &= \int p^\alpha(x) q^{1-\alpha}(x)\mathrm{d}x = \int p_F^{(\alpha)}(x;\theta_p) p_F^{1-\alpha}(x;\theta_q)\mathrm{d}x \\
&= \int \exp(\alpha(\langle x,\theta_p\rangle - F(\theta_p))) \times \exp((1-\alpha)(\langle x,\theta_q\rangle - F(\theta_q)))\mathrm{d}x \\
&= \int \exp\left(\langle x, \alpha\theta_p + (1-\alpha)\theta_q\rangle - (\alpha F(\theta_p) + (1-\alpha)F(\theta_q)\right)\mathrm{d}x \\
&= \exp-(\alpha F(\theta_p)+(1-\alpha)F(\theta_q)) \times \int \exp\left(\langle x,\alpha\theta_p+(1-\alpha)\theta_q\rangle - F(\alpha\theta_p+(1-\alpha)\theta_q) + F(\alpha\theta_p+(1-\alpha)\theta_q)\right)\mathrm{d}x \\
&= \exp\left(F(\alpha\theta_p+(1-\alpha)\theta_q) - (\alpha F(\theta_p)+(1-\alpha)F(\theta_q)\right) \times \int \exp\langle x,\alpha\theta_p+(1-\alpha)\theta_q\rangle - F(\alpha\theta_p+(1-\alpha)\theta_q)\mathrm{d}x \\
&= \exp\left(F(\alpha\theta_p+(1-\alpha)\theta_q) - (\alpha F(\theta_p)+(1-\alpha)F(\theta_q)\right) \times \underbrace{\int p_F(x;\alpha\theta_p+(1-\alpha)\theta_q)\mathrm{d}x}_{=1} \\
&= \exp(-\mathrm{BR}_F^{(\alpha)}(\theta_p,\theta_q)) \geq 0.
\end{aligned}$$

a *Bregman divergence* [14] (by taking the limit as $\alpha\rightarrow 1$ or $\alpha\rightarrow 0$).

Corollary 1: In the limit case $\alpha'\in\{0,1\}$, the α'-divergences amount to compute a Kullback-Leibler divergence, and is equivalent to compute a Bregman divergence for the log-normalized on the swapped natural parameters: $\mathrm{KL}(p_F(x;\theta_p),p_F(x;\theta_q)) = B_F(\theta_q,\theta_p)$.

Proof: The proof relies on the equivalence of Burbea-Rao divergences to Bregman divergences for extremal values of $\alpha\in\{0,1\}$.

$$\begin{aligned}
\mathrm{KL}(p,q) &= \mathrm{KL}(p_F(x;\theta_p),p_F(x;\theta_q)) && (65)\\
&= \lim_{\alpha'\rightarrow 1} D_{\alpha'}(p_F(x;\theta_p),p_F(x;\theta_q)) && (66)\\
&= \lim_{\alpha'\rightarrow 1}\frac{1}{\alpha'(1-\alpha')}(1-\underbrace{C_\alpha(p_F(x;\theta_p),p_F(x;\theta_q))}_{\text{since } \exp x\simeq_{x\simeq 0}1+x}) \\
&= \lim_{\alpha'\rightarrow 1}\frac{1}{\alpha'(1-\alpha')}\underbrace{\mathrm{BR}_F^{\alpha'}(\theta_p,\theta_q)}_{(1-\alpha')B_F(\theta_q,\theta_p)} && (67)\\
&= \lim_{\alpha'\rightarrow 1}\frac{1}{\alpha'}B_F(\theta_q,\theta_p) = B_F(\theta_q,\theta_p) && (68)
\end{aligned}$$

Similarly, we have $\lim_{\alpha'\rightarrow 0} D_{\alpha'}(p_F(x;\theta_p),p_F(x;\theta_q)) = \mathrm{KL}(p_F(x;\theta_q),p_F(x;\theta_p)) = B_F(\theta_p,\theta_q)$. ■

Table I reports the Bhattacharyya distances for members of the same exponential families.

C. Direct method for calculating the Bhattacharyya centroids of multivariate normals

To the best of our knowledge, the Bhattacharyya centroid has only been studied for univariate Gaussian or diagonal multivariate Gaussian distributions [42] in the context of speech recognition, where it is reported that it can be estimated using an iterative algorithm (no convergence guarantees are reported in [42]).

In order to compare this scheme on multivariate data with our generic Burbea-Rao scheme, we extend the approach of Rigazio et al. [42] to multivariate Gaussians. Plugging the Bhattacharyya distance of Gaussians in the energy function of the optimization problem (OPT), we get

$$\begin{aligned}
L(c) &= \sum_{i=1}^n \frac{1}{8}(\mu_c-\mu_i)^T\left(\frac{\Sigma_c+\Sigma_i}{2}\right)^{-1}(\mu_c-\mu_i) \\
&+ \frac{1}{2}\log\left(\frac{\det\left(\frac{\Sigma_c+\Sigma_i}{2}\right)}{\sqrt{\det\Sigma_c\det\Sigma_i}}\right). && (69)
\end{aligned}$$

This is equivalent to minimize the following energy:

$$\begin{aligned}
F(c) &= \sum_{i=1}^n (\mu_c-\mu_i)^T(\Sigma_c+\Sigma_i)^{-1}(\mu_c-\mu_i) \\
&+ 2\log\left(\det(\Sigma_c+\Sigma_i)\right) - \log\left(\det\Sigma_c\right) \\
&- \log\left(2^{2d}\det\Sigma_i\right). && (70)
\end{aligned}$$

In order to minimize $F(c)$, let us differentiate with respect to μ_c. let U_i denote $(\Sigma_c+\Sigma_i)^{-1}$. Using matrix differentials [43] (p.10 Eq. 73), we get:

$$\frac{\partial L}{\partial\mu_c} = \sum_{i=1}^n\left[U_i+U_i^T\right]\left[\mu_c-\mu_i\right] \tag{71}$$

Then one can estimate iteratively μ_c, since U_i depends on Σ_c which is unknown. We update μ_c as follows:

$$\mu_c(t+1) = \left[\sum_{i=1}^n\left[U_i+U_i^T\right]\right]^{-1}\left[\sum_{i=1}^n\left[U_i+U_i^T\right]\mu_i\right] \tag{72}$$

Now let us estimate Σ_c. We used matrix differentials [43] (p.9 Eq. 55 for the first term, and Eq. 51 p.8 for the two others):

$$\begin{aligned}
\frac{\partial L}{\partial\Sigma_c} &= \sum_{i=1}^n -U_i^T(\mu_c-\mu_i)(\mu_c-\mu_i)^T U_i^T \\
&+ 2\sum_{i=1}^n U_i^T - \sum_{i=1}^n\Sigma_c^{-T}. && (73)
\end{aligned}$$

Exponential family	$\tau : \lambda \rightarrow \theta$	$F(\theta)$ (up to a constant)	Bhattacharyya/Burbea-Rao $\mathrm{BR}_F(\lambda_p, \lambda_q) = \mathrm{BR}_F(\tau(\lambda_p), \tau(\lambda_q))$
Multinomial	$(\log \frac{p_i}{p_d})_i$	$\log(1 + \sum_{i=1}^{d-1} \exp \theta_i)$	$-\ln \sum_{i=1}^{d} \sqrt{p_i q_i}$
Poisson	$\log \lambda$	$\exp \theta$	$\frac{1}{2}(\sqrt{\mu_p} - \sqrt{\mu_q})^2$
Gaussian	$(\theta_1 = \mu, \theta_2 = \sigma^2)$	$-\frac{\theta_1^2}{4\theta_2} + \frac{1}{2}\log(-\frac{\pi}{\theta_2})$	$\frac{1}{4}\frac{(\mu_p-\mu_q)^2}{\sigma_p^2+\sigma_q^2} + \frac{1}{2}\ln\frac{\sigma_p^2+\sigma_q^2}{2\sigma_p\sigma_q}$
Multivariate Gaussian	$(\theta = \Sigma^{-1}\mu, \Theta = \frac{1}{2}\Sigma^{-1})$	$\frac{1}{4}\mathrm{tr}(\Theta^{-1}\theta\theta^T) - \frac{1}{2}\log\det\Theta$	$\frac{1}{8}(\mu_p - \mu_q)^T \left(\frac{\Sigma_p+\Sigma_q}{2}\right)^{-1}(\mu_p - \mu_q) + \frac{1}{2}\ln\frac{\det\frac{\Sigma_p+\Sigma_q}{2}}{\det\Sigma_p \det\Sigma_q}$

TABLE I

CLOSED-FORM BHATTACHARYYA DISTANCES FOR SOME CLASSES OF EXPONENTIAL FAMILIES (EXPRESSED IN SOURCE PARAMETERS FOR EASE OF USE)).

Taken into account the fact that Σ_c is symmetric, differential calculus on symmetric matrices can be simply estimate:

$$\frac{dL}{d\Sigma_c} = \frac{\partial L}{\partial \Sigma_c} + \left[\frac{\partial L}{\partial \Sigma_c}\right]^T - \mathrm{diag}\left(\frac{\partial L}{\partial \Sigma_c}\right). \tag{74}$$

Thus, if one notes

$$A = \sum_{i=1}^{n} 2U_i^T - U_i^T\left(\mu_c - \mu_i\right)\left(\mu_c - \mu_i\right)^T U_i^T \tag{75}$$

and recalling that Σ_c is symmetric, one has to solve

$$n(2\Sigma_c^{-1} - \mathrm{diag}(\Sigma_c^{-1})) = A + A^T - \mathrm{diag}(A). \tag{76}$$

Let

$$B = A + A^T - \mathrm{diag}(A) \tag{77}$$

Then one can estimate Σ_c iteratively as follows:

$$\Sigma_c^{(k+1)} = 2n\left[(B^{(k)} + \mathrm{diag}(B^{(k)}))\right]^{-1} \tag{78}$$

Let us now compare the two generic Burbea-Rao/tailored Gaussian methods for computing the Bhattacharyya centroids on multvariate Gaussians.

D. Applications to mixture simplification in statistics

Simplifying Gaussian mixtures is important in many applications arising in signal processing [26]. Mixture simplification is also a crucial step when one wants to study the Riemannian geometry induced by the Rao distance with respect to the Fisher metric: The set of mixture models need to have the same number of components, so that we simplify source mixtures to get a set of Gaussian mixtures with prescribed size. We adapt the hierarchical clustering algorithm of Garcia et al. [26] by replacing the symmetrized Bregman centroid (namely, the Jeffreys-Bregman centroid) by the Bhattacharyya centroid. We consider the task of color image segmentation by learning a Gaussian mixture model for each image. Each image is represented as a set of $5D$ points (color RGB and position xy).

The first experimental results depicted in Figure 3 demonstrates the *qualitative stability* of the clustering performance. In particular, the hierarchical clustering with respect to the Bhattacharrya distance performs qualitatively much better on the last `colormap` image.[12]

The second experiment focuses on characterizing the numerical convergence of the generic Burbea-Rao method compared to the tailored Gaussian method. Since we presented two novel different schemes to compute the Bhattacharyya centroids of multivariate Gaussians, one wants to compare them, both in terms of stability and accuracy. Whenever the ratio of Bhattacharyya distance energy function between those estimated centroids is greater than 1%, we consider that one of the two estimation methods is beaten (namely, the method that gives the highest Bhattacharyya distance). Among the 760 centroids computed to generate Figures 3, 100% were correct with the Burbea-Rao approach, while only 87% were correct with the tailored multivariate Gaussian matrix optimization method. The average number of iterations to reach the 1% accuracy is 4.1 for the Burbea-Rao estimation algorithm, and 5.2 for the alternative method.

Thus we experimentally checked that the generic CCCP iterative Burbea-Rao algorithm described for computing the Bhattacharrya centroids always converge, and moreover beats another *ad-hoc* iterative method tailored for multivariate Gaussians.

VI. CONCLUDING REMARKS

In this paper, we have shown that the Bhattacharrya distance for distributions of the same statistical exponential families can be computed equivalently as a Burbea-Rao divergence on the corresponding natural parameters. Those results extend to skew Chernoff coefficients (and Amari α-divergences) and skew Bhattacharyya distances using the notion of skew Burbea-Rao divergences. We proved that (skew) Burbea-Rao centroids can be efficiently estimated using an iterative concave-convex procedure with guaranteed convergence. We have shown that extremally skewed Burbea-Rao divergences amount asymptotically to evaluate Bregman divergences. This work emphasizes on the attractiveness of exponential families in Statistics. Indeed, it turns out that for many statistical distances, one can evaluate them in closed-form. For sake of brevity, we have not mentioned the recent β-divergences and γ-divergences [44], although their distances on exponential families are again available in closed-form.

The differential Riemannian geometry induced by the class of such Jensen difference measures was studied by Burbea and Rao [18], [19] who built quadratic differential metrics on probability spaces using Jensen differences. The Jensen-Shannon divergence is also an instance of a broad class of divergences called the f-divergences. A f-divergence I_f is a

[12] See reference images and segmentation using Bregman centroids at http://www.informationgeometry.org/MEF/

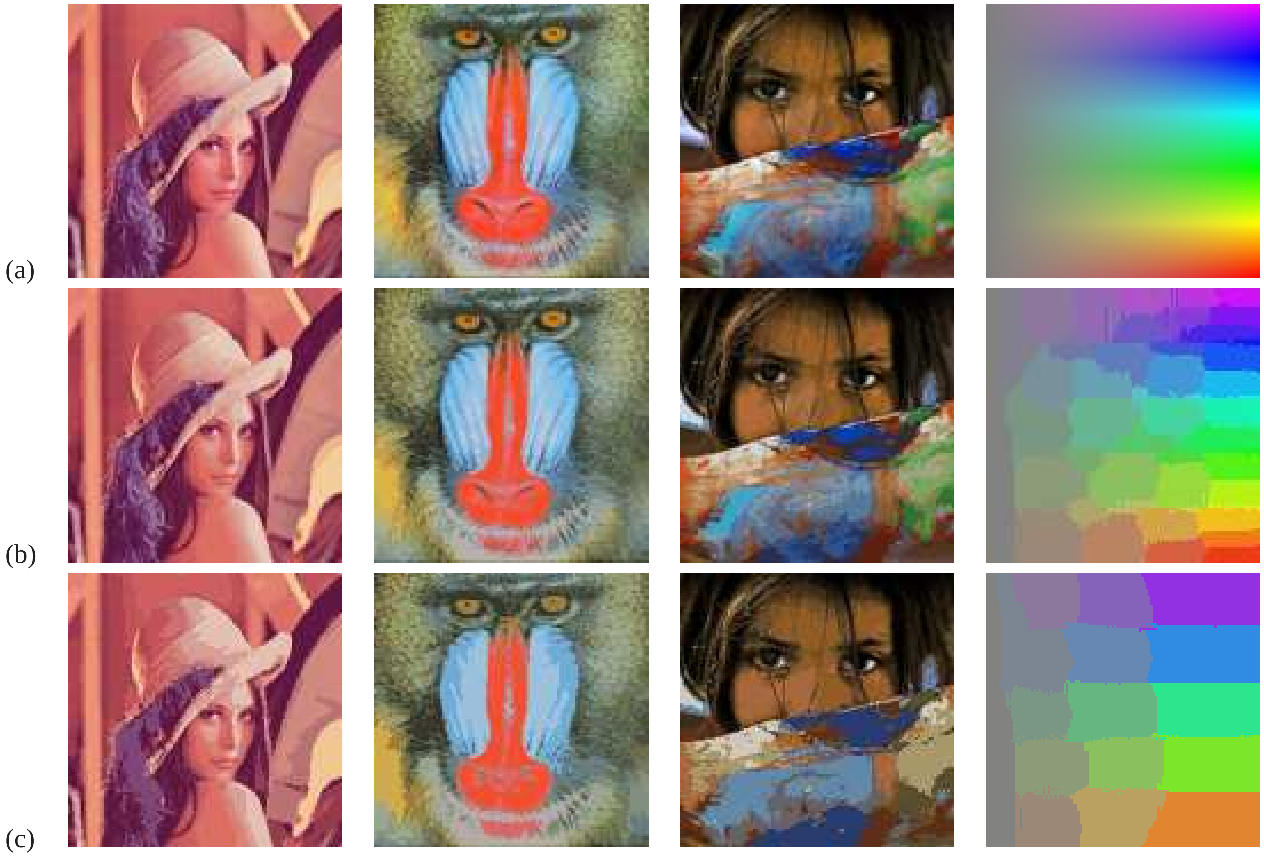

Fig. 3. Color image segmentation results: (a) source images, (b) segmentation with $k = 48$ 5D Gaussians, and (c) segmentation with $k = 16$ 5D Gaussians.

statistical measure of dissimilarity defined by the functional $I_f(p, q) = \int p(x) f(\frac{q(x)}{p(x)})\mathrm{d}x$. It turns out that the Jensen-Shannon divergence is a f-divergence for the generator

$$f(x) = \frac{1}{2}\left((x+1)\log\frac{2}{x+1} + x\log x\right). \qquad (79)$$

f-divergences preserve the information monotonicity [44], and their differential geometry was studied by Vos [45]. However, this Jensen-Shannon divergence is a very particular case of Burbea-Rao divergences since the squared Euclidean distance (another Burbea-Rao divergence) does not belong to the class of f-divergences.

Source code

The generic Burbea-Rao barycenter estimation algorithm shall be released in the JMEF open source library:

http://www.informationgeometry.org/MEF/

An applet visualizing the skew Burbea-Rao centroids ranging from the right-sided to left-sided Bregman centroids is available at: http://www.informationgeometry.org/BurbeaRao/

Acknowledgments

We gratefully acknowledge financial support from French agency DIGITEO (GAS 2008-16D) and French National Research Agency (ANR GAIA 07-BLAN-0328-01), and Sony Computer Science Laboratories, Inc. We are very grateful to the reviewers for their thorough and thoughtful comments and suggestions. In particular, we are thankful to the anonymous Referee that pointed out a rigorous proof of Lemma 1.

References

[1] A. N. Kolmogorov, "Sur la notion de la moyenne," *Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez.*, vol. 12, pp. 388–391, 1930.

[2] M. Nagumo, "Über eine Klasse der Mittelwerte," *Japanese Journal of Mathematics*, vol. 7, pp. 71–79, 1930, see Collected papers, Springer 1993.

[3] J. D. Aczél, "On mean values," *Bulletin of the American Mathematical Society*, vol. 54, no. 4, pp. 392–400, 1948, http://www.mta.hu/.

[4] A. Ben-Tal, A. Charnes, and M. Teboulle, "Entropic means," *Journal of Mathematical Analysis and Applications*, vol. 139, no. 2, pp. 537 – 551, 1989.

[5] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society, Series B*, vol. 28, pp. 131–142, 1966.

[6] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, p. 229318, 1967.

[7] F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," *IEEE Transactions on Information Theory*, vol. 55, no. 6, pp. 2048–2059, June 2009.

[8] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1997.

[9] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundational Trends in Machine Learning*, vol. 1, pp. 1–305, January 2008.

[10] S.-I. Amari, "α-divergence is unique, belonging to both f-divergence and bregman divergence classes," *IEEE Trans. Inf. Theor.*, vol. 55, no. 11, pp. 4925–4931, 2009.

[11] S.-i. Amari, "Integration of stochastic models by minimizing α-divergence," *Neural Comput.*, vol. 19, no. 10, pp. 2780–2796, 2007.

[12] S. Amari and H. Nagaoka, *Methods of Information Geometry*, A. M. Society, Ed. Oxford University Press, 2000.

[13] F. Nielsen and R. Nock, "The dual voronoi diagrams with respect to representational bregman divergences," in *International Symposium on Voronoi Diagrams (ISVD)*. DTU Lyngby, Denmark: IEEE, June 2009.

[14] J.-D. Boissonnat, F. Nielsen, and R. Nock, "Bregman Voronoi diagrams," *Discrete & Computational Geometry*, 2010, accepted, extend ACM-SIAM SODA 2007.

[15] I. Csiszár, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol. 68, no. 1-2, pp. 161–185, 1995.

[16] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[17] M. Detyniecki, "Mathematical aggregation operators and their application to video querying," Ph.D. dissertation, 2000.

[18] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," *IEEE Transactions on Information Theory*, vol. 28, no. 3, pp. 489–495, 1982.

[19] ——, "On the convexity of higher order Jensen differences based on entropy functions," *IEEE Transactions on Information Theory*, vol. 28, no. 6, pp. 961–, 1982.

[20] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, pp. 145–151, 1991.

[21] M. Basseville and J.-F. Cardoso, "On entropies, divergences and mean values," in *Proceedings of the IEEE Workshop on Information Theory*, 1995.

[22] L. M. Bregman, "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*, vol. 7, pp. 200–217, 1967.

[23] F. Nielsen and V. Garcia, "Statistical exponential families: A digest with flash cards," 2009, arXiv.org:0911.4863.

[24] K. Fukunaga, *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., 1990.

[25] L. Rigazio, B. Tsakam, and J. C. Junqua, "An optimal bhattacharyya centroid algorithm for gaussian clustering with applications in automatic speech recognition," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 3, 2000, pp. 1599–1602 vol.3. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=861998

[26] V. Garcia, F. Nielsen, and R. Nock, "Hierarchical gaussian mixture model," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2010.

[27] P. Chen, Y. Chen, and M. Rao, "Metrics defined by Bregman divergences: Part I," *Commun. Math. Sci.*, vol. 6, pp. 9915–926, 2008.

[28] ——, "Metrics defined by Bregman divergences: Part II," *Commun. Math. Sci.*, vol. 6, pp. 927–948, 2008.

[29] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London*, vol. 186, no. 1007, pp. 453–461, March 1946.

[30] F. Liese and I. Vajda, "On Divergences and Informations in Statistics and Information Theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, October 2006.

[31] J. Zhang, "Divergence function, duality, and convex analysis," *Neural Computation*, vol. 16, no. 1, pp. 159–195, 2004.

[32] A. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, 2003.

[33] B. Sriperumbudur and G. Lanckriet, "On the convergence of the concave-convex procedure," in *Neural Information Processing Systems*, 2009.

[34] Y. He, A. B. Hamza, and H. Krim, "An information divergence measure for ISAR image registration," in *Automatic target recognition XI (SPIE)*, vol. 4379, 2001, pp. 199–208.

[35] A. O. Hero, B. Ma, O. Michel, and J. D. Gorman, "Alpha-divergence for classification, indexing and retrieval," Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, Tech. Rep. 328, July, 2001, presented at Joint Statistical Meeting.

[36] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of Calcutta Mathematical Society*, vol. 35, pp. 99–110, 1943.

[37] K. Matusita, "Decision rules based on the distance, for problems of fit, two samples, and estimation," *Annal of Mathematics and Statistics*, vol. 26, pp. 631–640, 1955.

[38] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.

[39] F. Aherne, N. Thacker, and P. Rockett, "The Bhattacharyya metric as an absolute similarity measure for frequency coded data," *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1998.

[40] E. D. Hellinger, "Die orthogonalinvarianten quadratischer formen von unendlich vielen variablen," 1907, thesis of the university of Göttingen.

[41] S. Kakutani, "On equivalence of infinite product measures," *Annals of Mathematics*, vol. 49, no. 214-224, 1948.

[42] L. Rigazio, B. Tsakam, and J. Junqua, "Optimal Bhattacharyya centroid algorithm for Gaussian clustering with applications in automatic speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, 2000, pp. 1599–1602.

[43] K. B. Petersen and M. S. Pedersen, *The Matrix Cookbook*. Technical University of Denmark, oct 2008. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?3274

[44] A. Cichocki and S. ichi Amari, "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, 2010, review submitted.

[45] P. Vos, "Geometry of f-divergence," *Annals of the Institute of Statistical Mathematics*, vol. 43, no. 3, pp. 515–537, 1991.

PLACE
PHOTO
HERE

Frank Nielsen received the BSc (1992) and MSc (1994) degrees from Ecole Normale Superieure (ENS Lyon, France). He prepared his PhD on adaptive computational geometry at INRIA Sophia-Antipolis (France) and defended it in 1996. As a civil servant of the University of Nice (France), he gave lectures at the engineering schools ESSI and ISIA (Ecole des Mines). In 1997, he served in the army as a scientific member in the computer science laboratory of Ecole Polytechnique. In 1998, he joined Sony Computer Science Laboratories Inc., Tokyo (Japan) where he is senior researcher. He became a professor of the CS Dept. of Ecole Polytechnique in 2008. His current research interests include geometry, vision, graphics, learning, and optimization. He is a senior ACM and senior IEEE member.

PLACE
PHOTO
HERE

Sylvain Boltz Sylvain Boltz received the M.S. degree and the Ph.D. degree in computer vision from the University of Nice-Sophia Antipolis, France, in 2004 and 2008, respectively. Since then, he has been a postdoctoral fellow at the VisionLab, University of California, Los Angeles and a LIX-Qualcomm postdoctoral fellow in Ecole Polytechnique, France. His research spans computer vision and image, video processing with a particular interest in applications of information theory and compressed sensing to these areas.