Pattern learning and recognition on statistical manifolds: An information-geometric review

Frank Nielsen[1]

Sony Computer Science Laboratories, Inc.
Tokyo, Japan
Frank.Nielsen@acm.org
www.informationgeometry.org

Abstract. We review the *information-geometric* framework for statistical pattern recognition: First, we explain the role of statistical similarity measures and distances in fundamental statistical pattern recognition problems. We then concisely review the main statistical distances and report a novel versatile family of divergences. Depending on their intrinsic complexity, the statistical patterns are learned by either atomic parametric distributions, semi-parametric finite mixtures, or non-parametric kernel density distributions. Those statistical patterns are interpreted and handled geometrically in *statistical manifolds* either as single points, weighted sparse point sets or non-weighted dense point sets. We explain the construction of the two prominent families of statistical manifolds: The Rao Riemannian manifolds with geodesic metric distances, and the Amari-Chentsov manifolds with dual asymmetric non-metric divergences. For the latter manifolds, when considering atomic distributions from the same exponential families (including the ubiquitous Gaussian and multinomial families), we end up with dually flat exponential family manifolds that play a crucial role in many applications. We compare the advantages and disadvantages of these two approaches from the algorithmic point of view. Finally, we conclude with further perspectives on how "geometric thinking" may spur novel pattern modeling and processing paradigms.

Keywords: Statistical manifolds, mixture modeling, kernel density estimator, exponential families, clustering, Voronoi diagrams.

1 Introduction

1.1 Learning statistical patterns and the Cramér-Rao lower bound

Statistical pattern recognition [1] is concerned with *learning* patterns from observations using sensors, and with *analyzing* and *recognizing* those patterns efficiently. We shall consider three kinds of statistical models for learning patterns depending on their intrinsic complexities:

1. *parametric* models: A pattern is an atomic parametric distribution,

2. *semi-parametric models*: A pattern is a finite mixture of parametric distributions, and
3. *non-parametric models*: A pattern is a kernel density distribution.

Given a set of n observations $\{x_1, ..., x_n\}$, we may estimate the pattern parameter λ of the atomic distribution $p(x;\lambda)$ by using the *maximum likelihood principle*. The maximum likelihood estimator (MLE) proceeds by defining a function $L(\lambda; x_1, ..., x_n)$, called the *likelihood function* and maximizes this function with respect to λ. Since the sample is usually assumed to be *identically and independently distributed* (iid.), we have:

$$L(\lambda; x_1, ..., x_n) = \prod_i p(x_i; \lambda).$$

This maximization is equivalent (but mathematically often more convenient) to maximize the *log-likelihood function*:

$$l(\lambda; x_1, ..., x_n) = \log L(\lambda; x_1, ..., x_n) = \sum_i \log p(x_i; \lambda).$$

This maximization problem amounts to set the gradient to zero: $\nabla l(\lambda; x_1, ..., x_n) = 0$, and solve for the estimated quantity $\hat{\lambda}$ provided that it is well-defined (ie., that ML does not diverge to ∞). We can view the MLE as a function $\hat{\lambda}(X_1, ..., X_n)$ on a *random vector* and ask for its statistical performance. (Indeed, we can build a family of moment estimators by matching the sample l-th moments with the distribution l-th moments. This raises the question to compare them by analyzing, say, their variance characteristics.) Cramér [2], Fréchet [3] and Rao [4] independently proved a lower bound on the variance of any *unbiased* estimator $\hat{\lambda}$:

$$V[\hat{\lambda}] \succeq I(\lambda)^{-1},$$

where $\succeq$ denotes the Löwner partial ordering[1] on positive semidefinite matrices, and matrix $I(\lambda)$ is called the *Fisher information* matrix:

$$I(\lambda) = [I_{ij}(\lambda)], \quad I_{ij}(\lambda) = E[\partial_i l(x; \lambda) \partial_j l(x; \lambda)],$$

with ∂_k the shortcut notation: $\partial_k = \frac{\partial}{\partial \lambda_k}$. The Fisher information matrix [5] (FIM) is the variance of the score function $s(\lambda) = \nabla_\lambda \log p(\lambda; x)$: $I(\lambda) = V[s(\lambda)]$. This lower bound holds under very mild regularity conditions.

Learning finite mixtures of k atomic distributions is traditionally done using the *Expectation-Maximization* algorithm [6]. Learning a non-parametric distribution using a *kernel density estimator* (KDE) proceeds by choosing a kernel (e.g., Gaussian kernel), and by then fitting a kernel at each sample observation

[1] A symmetric matrix X is positive definite if and only if $\forall x \neq 0, x^\top X x > 0$, and $A \succeq B$ iff. $A - B \succ 0$. When the inequality is relaxed to include equality, we have the semi-positive definiteness property.

(controlling adaptively the kernel window is important in practice). Those three ML estimation/EM/KDE algorithms will be explained using the framework of information geometry in Section 5 when considering dually flat statistical *exponential family manifolds* (EFMs).

We now describe briefly the fundamental tasks of pattern recognition using eiher the *unsupervised setting* or the *supervised setting*. We recommend the introductory textbook [7] of Fukunaga for further explanations.

1.2 Unsupervised pattern recognition

Given a collection of n statistical patterns represented by their distributions (or estimated parameters $\lambda_1, ..., \lambda_n$), we would like to *categorize* them. That is, to identify groups (or clusters) of patterns inducing pattern categories. This is typically done using *clustering* algorithms. Observe that since patterns are represented by probability distributions, we need to have clustering algorithms suited to statistical distributions: Namely, clustering algorithms tailored for *information spaces*. We shall explain and describe the notions of statistical distances in information spaces in the following Section.

1.3 Supervised pattern recognition

When we are given beforehand a *training set* of properly *labeled* (or annotated) patterns, and seek to classify incoming online patterns, we may choose to label that *query pattern* with the label of its most similar annotated pattern in the database, or to vote by considering the k "nearest" patterns. Again, this requires a notion of statistical similarity that is described in Section 2.

1.4 Core geometric structures and algorithmic toolboxes

Since we are going to focus on two types of construction for defining *statistical manifolds of patterns*, let us review the wish list tools required by supervised or unsupervised pattern recognition. We need among others:

- Clustering (e.g., hard clustering à la k-means) with respect to statistical distances for unsupervised category discovery,
- To study the statistical Voronoi diagrams induced by the distinct category patterns,
- Data-structures for performing efficiently k-NN (nearest neighbor) search with respect to statistical distances (say, ball trees [8] or vantage point trees [9]),
- To study minimum enclosing balls (MEB) [10,11,12,13] (with applications in machine learning using vector ball machines [14])
- Etc.

1.5 Outline of the paper

The paper is organized as follows: In Section 2, we review the main statistical divergences, starting from the seminal Kullback-Leibler divergence, and explain why and how the intractable *distribution intersection similarity* measure needs to be upper bounded. This allows to explain the genesis of the Bhattacharyya divergence, the Chernoff information and the family of α-divergences. Following this interpretation, we further present the novel concept of *quasi-arithmetic α-divergences* and *quasi-arithmetic Chernoff informations*. Section 3 recalls that geometry is grounded by notion of invariance, and introduces the concepts of statistical invariance with the class of Ali-Silvey-Csiszár f-divergences [15,16]. We then describe two classical statistical manifold constructions: In Section 4, we present the *Rao Riemannian manifold* and discuss on its algorithmic considerations. In Section 5, we describe the *dual affine Amari-Chentsov manifolds*, and explain the process of learning parametric/semi-parametric/non-parametric patterns on those manifolds. Finally, Section 6 wrap ups this review paper and hints at further perspectives in the realm of statistical pattern analysis and recognition.

2 Statistical distances and divergences

2.1 The fundamental Kullback-Leibler divergence

The Kullback-Leibler divergence between two probability distributions $P(x)$ and $Q(x)$ (with density $p(x)$ and $q(x)$ with respect to a measure ν) is equal to the cross-entropy $H^\times(P:Q)$ minus the Shannon entropy $H(P)$:

$$\mathrm{KL}(P:Q)=\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\nu(x)=H^\times(P:Q)-H(P)\geq 0,$$

with

$$H^\times(P:Q)=\int -p(x)\log q(x)\mathrm{d}\nu(x),$$

$$H(P)=\int -p(x)\log p(x)\mathrm{d}\nu(x)=H^\times(P:P).$$

In practice, the Kullback-Leibler divergence $\mathrm{KL}(\tilde{P}:P)$ [17] can be interpreted as the distance between the *estimated distribution* $\tilde{P}$ (derived from the observed samples) and the *true hidden* distribution P. The Kullback-Leibler divergence does not satisfy the metric axioms of symmetry and triangular inequality. Therefore we call this dissimilarity[2] measure a *divergence* as it is a smooth and differentiable distance function that satisfies the essential separability property: $\mathrm{KL}(P:Q)=0$ if and only if $P=Q$. Computing the Kullback-Leibler may not be tractable analytically (eg., for patterns modeled by mixtures or KDEs)

[2] Note that there are Finslerian distances [34] that preserve the triangular inequality without being symmetric.

and requires costly Monte-Carlo stochastic approximation algorithms to estimate. To bypass this computational obstacle, several alternative distances like the Cauchy-Schwarz divergences [18] have been proposed. Since the inception of the Kullback-Leibler divergence, many other statistical distances have been proposed. We shall review in the context of classification the most prominent divergences.

2.2 Genesis of statistical distances

How can we define a notion of "distance" between two probability distributions P_1 and P_2 sharing the same support $\mathcal{X}$ with respective density p_1 and p_2 with respect to a dominating measure ν? What is the meaning of defining statistical distances? A distance $D(\cdot,\cdot)$ can be understood as a non-negative *dissimilarity measure* $D(P_1,P_2)\geq 0$ that is related to the notion of a *similarity measure* $0<S(P_1,P_2)\leq 1$. We present an overview of statistical distances based on the framework of *Bayesian binary hypothesis testing* [7].

Consider *discriminating* P_1 and P_2 with the following classification problem based on the mixture $P=\frac{1}{2}P_1+\frac{1}{2}P_2$. To sample mixture P, we first toss an unbiased coin and choose to sample from P_1 if the coin fell on heads or to sample from P_2 if it fell on tails. Thus mixture sampling is a *doubly stochastic process.* Now, given a *random variate* x of P (i.e., an observation) we would like to *decide* whether x was sampled from P_1 or from P_2? It makes sense to label x as class C_1 if $p_1(x)>p_2(x)$ and as class C_2, otherwise (if $p_2(x)\geq p_1(x)$). Since the distribution supports of P_1 and P_2 coincide, we can *never* be certain, and shall find a decision rule to minimize the risk. We seek for the best decision rule that minimizes the *probability of error* P_e, that is, the probability of misclassification. Consider the decision rule based on the *log-likelihood ratio* $\log\frac{p_1(x)}{p_2(x)}$:

$$\log\frac{p_1(x)}{p_2(x)}\underset{C_1}{\overset{C_2}{\lesseqgtr}}0.$$

The expected probability of error is:

$$P_e=E_P[\mathrm{error}(x)]=\int_{x\in\mathcal{X}}\mathrm{error}(x)p(x)\mathrm{d}\nu(x),$$

where $p(x)=\frac{1}{2}p_1(x)+\frac{1}{2}p_2(x)$ denotes the mixture density, and

$$\mathrm{error}(x)=\min\left(\frac{1}{2}\frac{p_1(x)}{p(x)},\frac{1}{2}\frac{p_2(x)}{p(x)}\right).$$

Indeed, suppose that at x (with probability $\frac{1}{2}$), $p_1(x)<p_2(x)$. Since we label x as C_2 then we misclassify with proportion $\frac{p_1(x)}{p(x)}$, and vice-versa [7]. Thus the probability of error $P_e=\frac{1}{2}S(P_1,P_2)$ where:

$$S(P_1,P_2)=\int\min(p_1(x),p_2(x))\mathrm{d}\nu(x).$$

S is a similarity measure since $S(P_1, P_2) = 1$ if and only if $P_1 = P_2$. It is known in computer vision, in the discrete case, as the *histogram intersection* similarity [19].

In practice, computing S is not tractable[3], specially for multivariate distributions. Thus, we seek to *upper bound* S using mathematically convenient tricks purposely designed for large classes of probability distributions. Consider the case of exponential families [20] that includes most common distributions such as Poisson, Gaussian, Gamma, Beta, Dirichlet, etc. distributions. Their natural canonical density decomposition is:

$$p_i = p(x|\theta_i) = \exp(\langle \theta_i, t(x) \rangle - F(\theta_i) + k(x)),$$

where θ_i is the natural parameter belonging to natural parameter space Θ. Function F is strictly convex and characterize the family. $t(x)$ is the sufficient statistic and $k(x)$ is an auxiliary carrier term [20]. Table 1 summarizes the canonical decomposition and related results for the multinomial and Gaussian families, with $p_i = p(x|\lambda_i) = p(x|\theta(\lambda_i))$. We can upper bound the probability intersection similarity S using the fact that:

$$\min(p_1(x), p_2(x)) \leq \sqrt{p_1(x) p_2(x)}.$$

We get:

$$S(P_1, P_2) \leq \rho(P_1, P_2) = \int \sqrt{p_1(x) p_2(x)} \mathrm{d}\nu(x).$$

The right hand-side is called the *Bhattacharrya coefficient* or *Bhattacharrya affinity.* For distributions belonging to the same exponential family (e.g., P_1 and P_2 are multivariate Gaussians [20]), we have:

$$\rho(P_1, P_2) = e^{-J_F(\theta_1, \theta_2)},$$

where J_F is a *Jensen divergence* defined over the natural parameter space:

$$J_F(\theta_1, \theta_2) = \frac{F(\theta_1) + F(\theta_2)}{2} - F\left(\frac{\theta_1 + \theta_2}{2}\right) \geq 0.$$

Of course, the bound is not the tightest. Therefore, we may consider for $\alpha \in (0, 1)$ that $\min(p_1(x), p_2(x)) \leq p_1(x)^\alpha p_2(x)^{1-\alpha}$. It follows the *$\alpha$-skewed Bhattacharrya coefficient* upper bounding S:

$$S(P_1, P_2) \leq \rho_\alpha(P_1, P_2) = \int p_1(x)^\alpha p_2(x)^{1-\alpha} \mathrm{d}\nu(x).$$

[3] In fact, using the mathematical rewriting trick $\min(a, b) = \frac{a+b}{2} - \frac{1}{2}|b - a|$, the probability intersection similarity is related to computing the *total variation metric distance*: $S(P_1, P_2) = 1 - \mathrm{TV}(P_1, P_2)$, with $\mathrm{TV}(P_1, P_2) = \frac{1}{2} \int |p_1(x) - p_2(x)| \mathrm{d}\nu(x)$. Bayes error that relies on a *cost design matrix* [7] to account for the different correct/incorrect classification costs extends the concept of the probability of error. Similarly, Bayes error can also be expressed using total variation distance on scaled probabilities (with scales depending on the prior mixture weights and on the cost design matrix).

This definition of affinity coefficient is still mathematically convenient for exponential families since we find that [20]:

$$\rho_\alpha(P_1,P_2)=e^{-J_F^{(\alpha)}(\theta_1,\theta_2)},$$

where $J_F^{(\alpha)}$ denotes a *skewed Jensen divergence* defined on the corresponding natural parameters:

$$J_F^{(\alpha)}(\theta_1,\theta_2)=\alpha F(\theta_1)+(1-\alpha)F(\theta_2)-F(\alpha\theta_1+(1-\alpha)\theta_2)\geq 0,$$

with equality to zero if and only if $\theta_1=\theta_2$ since F is a strictly convex and differentiable function. Setting $\alpha=\frac{1}{2}$, we get back the Bhattacharrya coefficient.

The upper bound can thus be "best" improved by optimizing over the α-range in $(0,1)$:

$$S(P_1,P_2)\leq\min_{\alpha\in[0,1]}\rho_\alpha(P_1,P_2)=\rho_{\alpha^*}(P_1,P_2)$$

The optimal value α^* is called *best error exponent* in Bayesian hypothesis testing [7]. For an iid. sequence of n observations, the probability of error is thus bounded [21] by:

$$P_e^{(n)}\leq\frac{1}{2}\rho_{\alpha^*}^n(P_1,P_2)$$

Historically, those similarity or affinity coefficients upper bounding the probability intersection similarity yielded respective notions of *statistical distances*:

$$B_\alpha(P_1,P_2)=-\log\rho_\alpha(P_1,P_2)=J_F^{(\alpha)}(\theta_1,\theta_2),$$

the *skew Bhattacharyya divergences*. Let us rescale B_α by a factor $\frac{1}{\alpha(1-\alpha)}$, then we have for $\alpha\not\in\{0,1\}$:

$$B'_\alpha(P_1,P_2)=\frac{1}{\alpha(1-\alpha)}B_\alpha(P_1,P_2)=\frac{1}{\alpha(1-\alpha)}J_F^{(\alpha)}(\theta_1,\theta_2)={J'}_F^{(\alpha)}(\theta_1,\theta_2).$$

When $\alpha\rightarrow 1$ or $\alpha\rightarrow 0$, we have B'_α that tends to the direct or reverse Kullback-Leibler divergence. For exponential families, that means that the scaled skew Jensen divergences ${J'}_F^{(\alpha)}$ tends to the direct or reverse Bregman divergence [22]:

$$\lim_{\alpha\rightarrow 0}{J'_F}^{(\alpha)}(\theta_1,\theta_2)=B_F(\theta_1,\theta_2),$$

where a Bregman divergence is defined for a strictly convex and differentiable genetor F by:

$$B_F(\theta_1,\theta_2)=F(\theta_1)-F(\theta_2)-(\theta_1-\theta_2)^\top\nabla F(\theta_2).$$

Furthermore, the *Chernoff divergence* (historically called *Chernoff information*) is defined by:

$$C(P_1, P_2) = \max_{\alpha \in [0,1]} -\log \rho_\alpha(P_1, P_2) = B_{\alpha^*}(P_1, P_2)$$

The mapping of a similarity coefficient by the monotonous function $-\log(\cdot)$ mimicked the unbounded property of the Kullback-Leibler divergence. However, we can also map a similarity coefficient $S \in (0, 1]$ to a distance $D \in [0, 1)$ by simply defining:

$$D(P_1, P_2) = 1 - S(P_1, P_2)$$

For example, we can define $d_\alpha(P_1, P_2) = 1 - \rho_\alpha(P_1, P_2)$. Since distances are used relatively to compare distributions and rank them as nearer or farther away, we can also rescale them. Another mathematical convenience is to scale d_α by $\frac{1}{\alpha(1-\alpha)}$ so that we get:

$$D_\alpha(P_1, P_2) = \frac{1 - \rho_\alpha(P_1, P_2)}{\alpha(1-\alpha)} = \frac{1 - \int p(x)^\alpha q(x)^{1-\alpha} \mathrm{d}\nu(x)}{\alpha(1-\alpha)}$$

This is known as the α-divergences of Amari that are the canonical divergences in information geometry [23]. When $\alpha \to 1$, we get the Kullback-Leibler divergence. When $\alpha \to 0$, we get the reverse Kullback-Leibler divergence. When $\alpha = \frac{1}{2}$, we find the (scaled) squared of the Hellinger distance. In information geometry, it is customary to set $\alpha \in [-\frac{1}{2}, \frac{1}{2}]$ instead of $[0, 1]$ by remapping $\alpha \leftarrow \alpha - \frac{1}{2}$. For members P_1 and P_2 belonging to the same exponential family, we have the following closed-formula for the α-divergences:

$$A_\alpha(P : Q) = \frac{4}{1-\alpha^2}\left(1 - \int_{x \in \mathcal{X}} p^{\frac{1-\alpha}{2}}(x) q^{\frac{1+\alpha}{2}} \mathrm{d}x\right),$$

$$A_\alpha(P : Q) = \frac{4}{1-\alpha^2}\left(1 - e^{-J_F^{\left(\frac{1-\alpha}{2}\right)}}(\theta(P) : \theta(Q))\right).$$

2.3 Novel quasi-arithmetic α-divergences and Chernoff information

Note that we can design many similar divergences by similarly upper bounding the probability intersection histogram similarity S. By definition, a *weighted mean* should have the property that it lies inside the range of its elements. Thus we can bound $\min(a, b)$ by *any* other kind of weighted means:

$$\min(a, b) \leq M(a, b; \alpha),$$

with $\alpha \in [0, 1]$. Instead of bounding S by a geometric weighted mean, let us consider for a strictly monotonous function f the *quasi-arithmetic weighted means*:

$$M_f(a, b; \alpha) = f^{-1}(\alpha f(a) + (1-\alpha) f(b)).$$

We get:

$$S(P_1,P_2) \leq \rho_\alpha^{(f)}(P_1,P_2) = \int M_f(p_1(x),p_2(x);\alpha)\mathrm{d}\nu(x),$$

for $\alpha \in (0,1)$, since the extremities $\alpha = 0,1$ are not discriminative:

$$\rho_0^{(f)}(P_1,P_2) = \rho_1^{(f)}(P_1,P_2) = 1.$$

When distributions coincide, notice that we have maximal affinity: $\rho_\alpha^{(f)}(P,P) = 1$.

Similarly, we can also generalize the Chernoff information to *quasi-arithmetic f-Chernoff information* as follows:

$$C_f(P_1,P_2) = \max_{\alpha\in[0,1]} -\log \int M_f(p_1(x),p_2(x))\mathrm{d}\nu(x).$$

For example, if we consider distributions not belonging to the exponential families like the univariate Cauchy distributions or the multivariate t-distributions (related to the unnormalized Pearson type VII elliptical distributions), in order to find a closed-form expression for $\int M_f(p_1(x),p_2(x))\mathrm{d}\nu(x)$, we may choose the *harmonic mean* with $f(x) = \frac{1}{x} = f^{-1}(x)$ instead of the geometric weighted mean.

To summarize, we have explained how the canonical α-divergences upper bounding the probability of error have been designed to include the sided (i.e., direct and reverse) Kullback-Leibler divergence, and explained the notion of probability separability using a binary classification task. We now turn our focus to build geometries for modeling statistical manifolds.

3 Divergence, invariance and geometry

In Euclidean geometry, we are familiar with the *invariant group* of *rigid transformations* (translations, rotations and reflections). The Euclidean distance $d(P_1,P_2)$ of two points P_1 and P_2 does not change if we apply such a rigid transformation T on their respective representations p_1 and p_2:

$$d(P_1,P_2) = d(p_1,p_2) = d(T(p_1),T(p_2)).$$

In fact, when we compute the distance between two points P_1 and P_2, we should not worry about the origin. Distance computations require numerical attributes that nevertheless should be invariant of the underlying geometry. Points exist beyond a specific coordinate system. This geometric invariance principle by a group of action has been carefully studied by Felix Klein in his *Erlangen* program.

A *divergence* is basically a smooth C_2 function (statistical distance) that may not be symmetric nor satisfy the triangular inequality of metrics. We denote by $D(P:Q)$ the divergence from distribution P (with density $p(x)$) to distribution Q (with density $q(x)$), where the ":" notation emphasizes the fact that this dissimilarity measure may not be symmetric: $D(P:Q) \neq D(Q:P)$.

	Multinomial ($n = 1$ trial)	Multivariate Gaussian
density	$\frac{1}{x_1!\cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$	$\frac{1}{(2\pi)^{d/2}\lvert\Sigma\rvert^{1/2}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right)$
support	$\{E_1, ..., E_k\}$	$\mathbb{R}^d$
base measure	Counting measure	Lebesgue measure
auxiliary carrier $k(x)$	$-\sum_{i=1}^{k} \log x_i!$	0
sufficient statistics $t(x)$	$(x_1, \cdots, x_{k-1})$	$(x, -xx^\top)$
$\theta(\lambda)$	$\left(\log\left(\frac{p_i}{p_k}\right)\right)_i$	$\left(\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}\right)$
Order D	$d-1$	$\frac{d}{d+3}2$
log-normalizer $F(\theta)$	$\log\left(1+\sum_{i=1}^{k-1}\exp\theta_i\right)$	$\frac{1}{4}\mathrm{tr}(\theta_2^{-1}\theta_1\theta_1^\top) - \frac{1}{2}\log\lvert\theta_2\rvert + \frac{d}{2}\log\pi$
$\eta = \nabla F(\theta) = E[t(X)]$	$\left(\frac{\exp\theta_i}{1+\sum_{j=1}^{k-1}\exp\theta_j}\right)_i$	$\left(\frac{1}{2}\theta_2^{-1}\theta_1, -\frac{1}{2}\theta_2^{-1} - \frac{1}{4}(\theta_2^{-1}\theta_1)(\theta_2^{-1}\theta_1)^\top\right)$
$\theta = \nabla F^*(\eta)$	$\left(\log\left(\frac{\eta_i}{1-\sum_{j=1}^{k-1}\eta_j}\right)\right)_i$	$\left(-(\eta_2+\eta_1\eta_1^\top)^{-1}\eta, -\frac{1}{2}(\eta_2+\eta_1\eta_1^\top)^{-1}\right)$
$F^*(\eta) = \langle\theta,\eta\rangle - F(\theta)$	$\left(\sum_{i=1}^{k-1}\eta_i\log\eta_i\right) + \left(1-\sum_{i=1}^{k-1}\eta_i\right)\log\left(1-\sum_{i=1}^{k-1}\eta_i\right)$	$-\frac{1}{2}\log\left(1+\eta_1^\top\eta_2^{-1}\eta_1\right) - \frac{1}{2}\log\lvert-\eta_2\rvert - \frac{d}{2}\log(2\pi e)$
Kullback-Leibler divergence	$p_{1,k}\log\frac{p_{1,k}}{p_{2,k}} - \sum_{i=1}^{k-1} p_{1,i}\log\frac{p_{2,i}}{p_{1,i}}$	$\frac{1}{2}\left(\log\left(\frac{\lvert\Sigma_2\rvert}{\lvert\Sigma_1\rvert}\right) + \mathrm{tr}\left(\Sigma_2^{-1}\Sigma_1\right)\right) + \frac{1}{2}\left((\mu_2-\mu_1)^\top\Sigma_2^{-1}(\mu_2-\mu_1) - d\right)$
Fisher information $I(\theta) = \nabla^2 F(\theta)$	$I_{ii} = \frac{1}{p_i} + \frac{1}{p_d}, \quad I_{ij} = \frac{1}{p_d}(i \neq j)$	$I_{ij} = \partial_i\mu^\top\Sigma^{-1}\partial_j\mu$
MLE $\hat{\eta} = \bar{t}$	$\hat{p}_i = \frac{n_i}{n}$	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})\top$
$\lambda(\theta)$	$\begin{cases} p_i = \frac{\exp\theta_i}{1+\sum_{j=1}^{k-1}\exp\theta_j} & \text{if } i < k \\ p_k = \frac{1}{1+\sum_{j=1}^{k-1}\exp\theta_j} \end{cases}$	$\left(\frac{1}{2}\theta_2^{-1}\theta, \frac{1}{2}\theta_2^{-1}\right)$
$\lambda(\eta)$	$\begin{cases} p_i = \eta_i & \text{if } i < k \\ p_k = 1 - \sum_{j=1}^{k-1}\eta_j \end{cases}$	$\left(\eta, -(\eta_2+\eta\eta^\top)\right)$

Table 1. Summary of the canonical decompositions and its related results for the two prominent multivariate exponential families [20] met in statistical pattern recognition: The multinomial (one trial) and the Gaussian distributions. Observe that both families have their log-normalizer F *not* separable. For the MLE, we have that $\sqrt{n}(\hat{\theta}-\theta)$ that converges in distribution to $N(0, I^{-1}(\theta))$.

It is proven that the only *statistical invariant* divergences [23,24] are the Ali-Silvey-Csiszár f-divergences D_f [15,16] that are defined for a functional convex generator f satisfying $f(1) = f'(1) = 0$ and $f''(1) = 1$ by:

$$D_f(P : Q) = \int_{x\in\mathcal{X}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\nu(x).$$

Indeed, under an invertible mapping function (with $\dim(\mathcal{X}) = \dim(\mathcal{Y}) = d$):

$$\begin{aligned} m : \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto y = m(x) \end{aligned}$$

a probability density $p(x)$ is converted into another probability density $q(y)$ such that:

$$p(x)\mathrm{d}x = q(y)\mathrm{d}y, \qquad \mathrm{d}y = |M(x)|\mathrm{d}x,$$

where $|M(x)|$ denotes the determinant of the Jacobian matrix [23] of the transformation m (i.e., the partial derivatives):

$$M(x) = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_d}{\partial x_1} & \cdots & \frac{\partial y_d}{\partial x_d} \end{bmatrix}.$$

It follows that we have:

$$q(y) = q(m(x)) = p(x)|M(x)|^{-1}.$$

For any two densities p_1 and p_2, we have the f-divergence on the transformed densities q_1 and q_2 that can be rewritten mathematically as:

$$\begin{aligned} D_f(q_1 : q_2) &= \int_{y\in\mathcal{Y}} q_1(y) f\left(\frac{q_2(y)}{q_1(y)}\right) \mathrm{d}y, \\ &= \int_{x\in\mathcal{X}} p_1(x)|M(x)|^{-1} f\left(\frac{p_2(x)}{p_1(x)}\right) |M(x)|\mathrm{d}x, \\ &= D_f(p_1 : p_2). \end{aligned}$$

Furthermore, the f-divergences are the only divergences satisfying the *data-processing theorem* [25]. This theorem characterizes the property of *information monotonicity* [26]. Consider discrete distributions on an alphabet $\mathcal{X}$ of d letters. For any partition $\mathcal{B} = \mathcal{X}_1 \cup ...\mathcal{X}_b$ of $\mathcal{X}$ that merge alphabet letters into $b \leq d$ bins, we have

$$0 \leq D_f(\bar{p}_1 : \bar{p}_2) \leq D_f(p_1 : p_2),$$

where $\bar{p}_1$ and $\bar{p}_2$ are the discrete distribution induced by the partition $\mathcal{B}$ on $\mathcal{X}$. That is, we loose discrimination power by coarse-graining the support of the distributions. The most fundamental f-divergence is the Kullback-Leibler divergence [17] obtained for the generator $f(x) = x \log x$: In general, statistical invariance is characterized under *Markov morphisms* [27,24] (also called *sufficient stochastic kernels* [24]) that generalizes the deterministic transformations $y = m(x)$. Loosely speaking, a geometric parametric statistical manifold $\mathcal{F} = \{p_\theta(x) | \theta \in \Theta\}$ equipped with a f-divergence must also provide invariance by:

Non-singular parameter re-parameterization. That is, if we choose a different coordinate system, say $\theta' = f(\theta)$ for an invertible transformation f, it should not impact the intrinsic distance between the underlying distributions. For example, whether we parametrize the Gaussian manifold by $\theta = (\mu, \sigma)$ or by $\theta' = (\mu^5, \sigma^4)$, it should preserve the distance.

Sufficient statistic. When making statistical inference, we use statistics $T : \mathbb{R}^d \rightarrow \Theta \subseteq \mathbb{R}^D$ (e.g., the mean statistic $T_n(X) = \frac{1}{n}\sum_{i=1}^n X_i$ is used for estimating the parameter μ of Gaussians). In statistics, the concept of *sufficiency* was introduced by Fisher [28]:
Mathematically, the fact that all information should be aggregated inside the sufficient statistic is written as

$$\Pr(x|t, \theta) = \Pr(x|t).$$

It is not surprising that all statistical information of a parametric distribution with D parameters can be recovered from a set of D statistics. For example, the univariate Gaussian with $d = \dim(\mathcal{X}) = 1$ and $D = \dim(\Theta) = 2$ (for parameters $\theta = (\mu, \sigma)$) is recovered from the mean and variance statistics. A sufficient statistic is a set of statistics that compress *information without loss* for statistical inference.

4 Rao statistical manifolds: A Riemannian approach

4.1 Riemannian construction of Rao manifolds

We review the construction first reported in 1945 by C.R. Rao [4]. Consider a family of parametric probability distribution $\{p_\theta(x)\}_\theta$ with $x \in \mathbb{R}^d$ (dimension of the support) and $\theta \in \mathbb{R}^D$ denoting the D-dimensional parameters of the distributions. It is called the order of the probability family. The *population parameter space* is defined by:

$$\Theta = \left\{ \theta \in \mathbb{R}^D \middle| \int p_\theta(x) \mathrm{d}x = 1 \right\}.$$

A given distribution $p_\theta(x)$ is interpreted as a corresponding point indexed by $\theta \in \mathbb{R}^D$. θ also encodes a coordinate system to identify probability models: $\theta \leftrightarrow p_\theta(x)$.

Consider now two infinitesimally close points θ and $\theta + \mathrm{d}\theta$. Their probability densities differ by their first order differentials: $\mathrm{d}p(\theta)$. The distribution of $\mathrm{d}p$ over all the support aggregates the consequences of replacing θ by $\theta + \mathrm{d}\theta$. Rao's revolutionary idea was to consider the *relative discrepancy* $\frac{\mathrm{d}p}{p}$ and to take the variance of this difference distribution to define the following *quadratic differential form*:

$$\begin{aligned}\mathrm{d}s^2(\theta) &= \sum_{i=1}^{D}\sum_{j=1}^{D} g_{ij}(\theta)\mathrm{d}\theta_i\mathrm{d}\theta_j,\\ &= (\nabla\theta)^\top G(\theta)\nabla\theta,\end{aligned}$$

with the matrix entries of $G(\theta) = [g_{ij}(\theta)]$ as

$$g_{ij}(\theta) = E_\theta\left[\frac{1}{p(\theta)}\frac{\partial p}{\partial\theta_i}\frac{1}{p(\theta)}\frac{\partial p}{\partial\theta_j}\right] = g_{ji}(\theta).$$

In differential geometry, we often use the symbol ∂_i as a shortcut to $\frac{\partial}{\partial\theta_i}$.

The elements $g_{ij}(\theta)$ form the quadratic differential form defining the elementary length of Riemannian geometry. The matrix $G(\theta) = [g_{ij}(\theta)] \succ 0$ is positive definite and turns out to be equivalent to the *Fisher information matrix*: $G(\theta) = I(\theta)$. The information matrix is invariant to monotonous transformations of the parameter space [4] and makes it a good candidate for a Riemannian metric as the concepts of the concepts of invariance in statistical manifolds[29,27] later was revealed.

4.2 Rao Riemannian geodesic metric distance

Let P_1 and P_2 be two points of the population space corresponding to the distributions with respective parameters θ_1 and θ_2. In Riemannian geometry, the geodesics are the *shortest paths*. The statistical distance between the two populations is defined by integrating the infinitesimal element lengths ds along the geodesic linking P_1 and P_2. Equipped with the Fisher information matrix tensor $I(\theta)$, the *Rao distance* $D(\cdot,\cdot)$ between two distributions on a statistical manifold can be calculated from the geodesic length as follows:

$$D(p_{\theta_1}(x), p_{\theta_2}(x)) = \min_{\substack{\theta(t)\\ \theta(0)=\theta_1,\theta(1)=\theta_2}} \int_0^1 \left(\sqrt{(\nabla\theta)^\top I(\theta)\nabla\theta}\right)\mathrm{d}t \qquad (1)$$

Therefore we need to calculate explicitly the geodesic linking $p_{\theta_1}(x)$ to $p_{\theta_2}(x)$ to compute Rao's distance. This is done by solving the following second order ordinary differential equation (ODE) [23]:

$$g_{ki}\ddot{\theta}_i + \mathit{\Gamma}_{k,ij}\dot{\theta}_i\dot{\theta}_j = 0,$$

where Einstein summation [23] convention has been used to simplify the mathematical writing by removing the leading sum symbols. The coefficients $\mathit{\Gamma}_{k,ij}$ are the Christoffel symbols of the first kind defined by:

$$\mathit{\Gamma}_{k,ij} = \frac{1}{2}\left(\frac{\partial g_{ik}}{\partial\theta_j} + \frac{\partial g_{kj}}{\partial\theta_i} - \frac{\partial g_{ij}}{\partial\theta_k}\right).$$

For a parametric statistical manifold with D parameters, there are D^3 Christoffel symbols. In practice, it is difficult to explicitly compute the geodesics of the Fisher-Rao geometry of arbitrary models, and one needs to perform a gradient descent to find a local solution for the geodesics [30]. This is a drawback of the Rao's distance as it has to be checked manually whether the integral admits a closed-form expression or not.

To give an example of the Rao distance, consider the smooth manifold of univariate normal distributions, indexed by the $\theta = (\mu, \sigma)$ coordinate system. The Fisher information matrix is

$$I(\theta) = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \succ 0. \tag{2}$$

The infinitesimal element length is:

$$\begin{aligned} \mathrm{d}s^2 &= (\nabla\theta)^\top I(\theta) \nabla\theta, \\ &= \frac{\mathrm{d}\mu^2}{\sigma^2} + \frac{2\mathrm{d}\sigma^2}{\sigma^2}. \end{aligned}$$

After the minimization of the path length integral, the Rao distance between two normal distributions [4,31] $\theta_1 = (\mu_1, \sigma_1)$ and $\theta_2 = (\mu_2, \sigma_2)$ is given by:

$$D(\theta_1, \theta_2) = \begin{cases} \sqrt{2} \log \frac{\sigma_2}{\sigma_1} & \text{if } \mu_1 = \mu_2, \\ \frac{|\mu_1 - \mu_2|}{\sigma} & \text{if } \sigma_1 = \sigma_2 = \sigma, \\ \sqrt{2} \log \frac{\tan \frac{a_1}{2}}{\tan \frac{a_2}{2}} & \text{otherwise.} \end{cases}$$

where $a_1 = \arcsin \frac{\sigma_1}{b_{12}}$, $a_2 = \arcsin \frac{\sigma_2}{b_{12}}$ and

$$b_{12} = \sigma_1^2 + \frac{(\mu_1 - \mu_2)^2 - 2(\sigma_2^2 - \sigma_1^2)}{8(\mu_1 - \mu_2)^2}.$$

For univariate normal distributions, Rao's distance amounts to computing the hyperbolic distance for $\mathbb{H}(\frac{1}{\sqrt{2}})$, see [32].

The table below summarizes some types of Rao geometries:

Riemannian geometry	Fisher-Rao statistical manifold
Euclidean	Normal distributions with same covariance matrices
Spherical	Discrete distributions (multinomials)
Hyperbolic	Location-scale family (i.e, univariate normal, Cauchy)

4.3 Geometric computing on Rao statistical manifolds

Observe that in any tangent plane T_x of the Rao statistical manifold, the inner product induces a squared Mahalanobis distance:

$$D_x(p, q) = (p - q)^\top I(x)(p - q).$$

Since matrix $I(x) \succ 0$ is positive definite, we can apply Cholesky decomposition on the Fisher information matrix $I(x) = L(x)L^\top(x)$, where $L(x)$ is a lower triangular matrix with strictly positive diagonal entries. By mapping the points p to $L(p)^\top$ in the tangent space T_p, the squared Mahalanobis amounts to computing the squared Euclidean distance $D_E(p,q) = \|p-q\|^2$ in the tangent planes:

$$D_x(p,q) = (p-q)^\top I(x)(p-q) = (p-q)^\top L(x)L^\top(x)(p-q) = D_E(L^\top(x)p, L^\top(x)q).$$

It follows that after applying the "Cholesky transformation" of objects into the tangent planes, we can solve geometric problems in tangent planes as one usually does in the Euclidean geometry. Thus we can use the classic toolbox of computational geometry in tangent planes (for extrinsic computing and mapping back and forth on the manifold using the Riemannian Log/Exp).

Let us consider the Rao univariate normal manifold that is equivalent to the hyperbolic plane. Classical algorithms like the clustering k-means do not apply straightforwardly because, in hyperpolic geometry, computing a center of mass e is not available in closed-form but requires a numerical scheme. To bypass this limitation, we rather consider non-Kärcher centroids called *model centroids* that can be easily built in hyperbolic geometry [33,34]. The computational geometry toolbox is rather limited even for the hyperbolic geometry. We proved that hyperbolic Voronoi diagrams is affine in the Klein model and reported an optimal algorithm based on power diagram construction [35,36]. We alo generalized the Euclidean minimum enclosing ball approximation algorithm using an iterative geodesic cut algorithm in [13]. This is useful for zero-centered multivariate normal distributions that has negative curvature and is guaranteed to converge.

In general, the algorithmic toolbox on generic Riemannian manifolds is very restricted due to the lack of closed-form expressions for the geodesics. One of the techniques consists in using the Riemannian Log/Exp mapping to go from/to the manifold to the tangent planes. See [37] for a review with applications on computational anatomy.

The next section explains the dual affine geometry induced by a convex function (with explicit dual geodesic parameterizations) and shows how to design efficient algorithms when consider the exponential family manifolds.

5 Amari-Chentsov statistical manifolds

5.1 Construction of dually flat statistical manifolds

The Legendre-Fenchel convex duality is at the core of information geometry: Any strictly convex and differentiable function F admits a dual convex conjugate F^* such that:

$$F^*(\eta) = \max_{\theta \in \Theta} \theta^\top \eta - F(\theta).$$

The maximum is attained for $\eta = \nabla F(\theta)$ and is unique since $F(\theta)$ is strictly convex ($\nabla^2 F(\theta) \succ 0$). It follows that $\theta = \nabla F^{-1}(\eta)$, where ∇F^{-1} denotes the functional inverse gradient. This implies that:

$$F^*(\eta) = \eta^\top (\nabla F)^{-1}(\eta) - F((\nabla F)^{-1}(\eta)).$$

The Legendre transformation is also called slope transformation since it maps $\theta \to \eta = \nabla F(\theta)$, where $\nabla F(\theta)$ is the gradient at θ, visualized as the slope of the support tangent plane of F at θ. The transformation is an involution for strictly convex and differentiable functions: $(F^*)^* = F$. It follows that gradient of convex conjugates are reciprocal to each other: $\nabla F^* = (\nabla F)^{-1}$. Legendre duality induces dual coordinate systems:

$$\begin{aligned} \eta &= \nabla F(\theta), \\ \theta &= \nabla F^*(\eta). \end{aligned}$$

Furthermore, those dual coordinate systems are orthogonal to each other since,

$$\nabla^2 F(\theta) \nabla^2 F^*(\eta) = \mathrm{Id},$$

the identity matrix.

The Bregman divergence can also be rewritten in a canonical mixed coordinate form C_F or in the θ- or η-coordinate systems as

$$\begin{aligned} B_F(\theta_2 : \theta_1) &= F(\theta_2) + F^*(\eta_1) - \theta_2^\top \eta_1 = C_F(\theta_2, \eta_1) = C_{F^*}(\eta_1, \theta_2), \\ &= B_{F^*}(\eta_1 : \eta_2). \end{aligned}$$

Another use of the Legendre duality is to interpret the log-density of an exponential family as a dual Bregman divergence [38]:

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

with $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$.

5.2 Dual geodesics: Exponential and mixture geodesics

Information geometry as further pioneered by Amari [23] considers dual affine geometries introduced by a pair of connections: the α-connection and $-\alpha$-connection instead of taking the Levi-Civita connection induced by the Fisher information Riemmanian metric of Rao. The ± 1-connections give rise to dually flat spaces [23] equipped with the Kullback-Leibler divergence [17]. The case of $\alpha = -1$ denotes the mixture family, and the exponential family is obtained for $\alpha = 1$. We omit technical details in this expository paper, but refer the reader to the monograph [23] for details.

For our purpose, let us say that the geodesics are defined not anymore as shortest path lengths (like in the metric case of the Fisher-Rao geometry) but rather as curves that ensures the parallel transport of vectors [23]. This defines the notion of “straightness” of lines. Riemannian geodesics satisfy both the straightness property and the minimum length requirements. Introducing dual connections, we do not have anymore distances interpreted as curve lengths, but the geodesics defined by the notion of straightness only.

In information geometry, we have dual geodesics that are expressed for the exponential family (induced by a convex function F) in the dual affine coordinate systems θ/η for $\alpha = \pm 1$ as:

$$\gamma_{12} : L(\theta_1, \theta_2) = \{\theta = (1-\lambda)\theta_1 + \lambda\theta_2 \mid \lambda \in [0,1]\},$$
$$\gamma_{12}^* : L^*(\eta_1, \eta_2) = \{\eta = (1-\lambda)\eta_1 + \lambda\eta_2 \mid \lambda \in [0,1]\}.$$

Furthermore, there is a *Pythagorean theorem* that allows one to define information-theoretic projections [23]. Consider three points p, q and r such that γ_{pq} is the θ-geodesic linking p to q, and γ_{qr}^* is the η-geodesic linking q to r. The geodesics are orthogonal at the intersection point q if and only if the Pythagorean relation is satisfied:

$$D(p:r) = D(p:q) + D(q:r).$$

In fact, a more general triangle relation (extending the law of cosines) exists:

$$D(p:q) + D(q:r) - D(p:r) = (\theta(p) - \theta(q))^\top (\eta(r) - \eta(q)).$$

Note that the θ-geodesic γ_{pq} and η-geodesic γ_{qr}^* are orthogonal with respect to the inner product $G(q)$ defined at q (with $G(q) = I(q)$ being the Fisher information matrix at q). Two vectors u and v in the tangent place T_q at q are said to be orthogonal if and only if their inner product equals zero:

$$u \perp_q v \Leftrightarrow u^\top I(q) v = 0.$$

Information geometry of dually flat spaces thus extend the traditional self-dual Euclidean geometry, obtained for the convex function $F(x) = \frac{1}{2}x^\top x$ (and corresponding to the statistical manifold of isotropic Gaussians).

The construction can be extended to dual constant curvature manifolds using Amari-Chentsov's affine α-connections. We omit those details here, but refer the reader to the textbook [23].

5.3 Learning statistical patterns

We mentioned in the introduction that statistical patterns can either be learned from (1) a parametric model, (2) a mixture model, or (3) a kernel density estimator. We concisely review algorithms to learn those statistical patterns by taking into consideration the *exponential family manifold* (EFM).

Parametric distribution Let $x_1, ..., x_n$ be n data points assumed to be iid. from an exponential family. The maximum likelihood estimator (MLE) yields [20]:

$$\eta(\hat{P}) = \frac{1}{n} t(x_i) = \bar{t}$$

The point $\hat{P}$ on the EFM with η-coordinates $\bar{t}$ is called the *observed point* in information geometry [23]. The MLE is guaranteed to exist [39,40] provided that matrix:

$$T = \begin{bmatrix} 1 \; t_1(x_1) \; ... \; t_D(x_1) \\ \vdots \; \vdots \qquad\quad \vdots \;\; \vdots \\ 1 \; t_1(x_n) \; ... \; t_D(x_n) \end{bmatrix} \tag{3}$$

of dimension $n \times (D+1)$ has rank $D+1$ [40].

Furthermore, the log-likelihood achieved by the MLE can be expressed as:

$$l(\hat{\theta}; x_1, ..., x_n) = F^*(\hat{\eta}) + \frac{1}{n} \sum_{i=1}^{n} k(x_i)$$

For exponential families, the MLE is consistent and efficient (i.e., matches the Cramér-Rao lower bound) and has normal asymptotic distribution with covariance matrix the inverse of the Fisher information matrix:

$$\sqrt{n}(\hat{\theta} - \theta) \overset{\text{distribution}}{\longrightarrow} N(0, I^{-1}(\theta)).$$

Notice that to choose between two different exponential family models, say, parameterized by F_1 and F_2, we can evaluate their MLE log-likelihood using their respective convex conjugates F_1^* and F_2^*, and choose the model which yielded the highest likelihood.

Learning finite mixture distributions By using the duality between (regular) exponential families and (regular) Bregman divergences, Banerjee *et al.* [38] showed that the classical EM algorithm for learning mixtures of the same exponential families amount to a *soft Bregman clustering*. The EM maximizes the expected complete log-likelihood [7]. Recently, it has been shown that maximizing the complete log-likelihood (by labeling all observation data with their component number) for an exponential family mixture amounts to perform a k-means clustering for the dual Bregman divergence B_{F^*} on the sufficient statistic data: $\{y_i = t(x_i)\}_{i=1}^n$. Thus by using Lloyd batched k-means algorithm that optimizes the k-means loss, we obtain an algorithm for learning mixtures. This algorithm is called k-MLE [41] and outperforms computationally EM since it deals with hard membership. Furthermore, a generalization of k-MLE considers for each component a different exponential family and adds a step to choose the best exponential family of a cluster. This generalized k-MLE has been described specifically for learning generalized gaussian mixtures [42], gamma mixtures [43], and Wishart mixtures [44]. (The technical details focus on computing the dual convex conjugate F^* and on how to stratify an exponential family with $D > 1$ parameters as a family of exponential families of order $D - 1$.)

Learning non-parametric distributions with KDEs For each datum x_i, we can associate a density with weight $\frac{1}{n}$ and mode matching x_i. This is the kernel density estimator [7] (KDE). For the kernel family, we can choose the univariate location-scale families or multivariate elliptical distributions. Normal distributions belong both to the exponential families and the elliptical families.

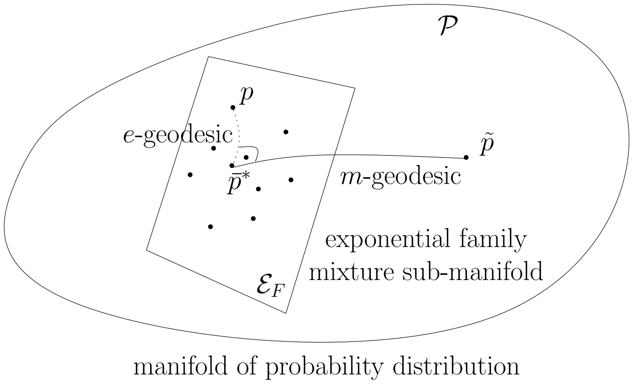

Fig. 1. Simplifying a statistical mixture of exponential families or KDE $\tilde{p}$ to a single component model amounts to perform a Kullback-Leibler projection of the mixture onto the exponential family manifold [45]. Optimality is proved using the Pythagorean theorem of dually flat geometries.

Since the mixture model is dense and has n components, we can simplify this representation to a sparse model by performing mixture simplification.

Simplifying KDEs and mixtures A statistical mixture or a KDE is represented on the exponential family manifold as a *weighted point set*. We simplify a mixture by clustering. This requires to compute centroids and barycenters with respect to information-theoretic distances. The Kullback-Leibler and Jeffreys centroid computations have been investigated in [46].

A neat geometric characterization of the mixture simplification is depicted in Figure 1. We project the mixture $\tilde{p}$ on the exponential family manifold using the m-geodesic. This amounts to compute a barycenter of the weighted parameter points on the manifold. See [45] for further details.

Instead of clustering groupwise, we can also consider hierarchical clustering to get a dendrogram [7] (a binary tree-structured representation): This yields a mixture representation with *levels of details* for modeling statistical mixtures [47]. We can extend the centroid computations to the wider class of skewed Bhattacharrya centroids [22] that encompasses the Kullback-Leibler divergence. In [48,49], we further consider the novel class of information-theoretic divergences called *total Bregman divergences*. The total Bregman divergence (and total Kullback-Leibler divergence when dealing with exponential family members) is defined by:

$$tB(P:Q) = \frac{B(P:Q)}{\sqrt{1+\|\nabla F(\theta(Q))\|^2}},$$

and yields *conformal geometry* [49]. We experimentally improved application performance for shape retrieval and diffusion tensor imaging.

5.4 Statistical Voronoi diagrams

It is well-known that the k-means algorithm [7] is related to ordinary Voronoi diagrams since data points are associated to their closest centroid. Namely, the centroids play the role of Voronoi seeds. The Kullback-Leibler k-means intervenes in the description of the k-MLE or the mixture simplification algorithms. For distributions belonging to the same exponential families, those statistical Voronoi diagrams amount to perform Bregman Voronoi diagrams on the distribution parameters (using either the natural θ-coordinates, or the dual η-coordinates). The Bregman Voronoi diagrams and its extensions have been investigated in [50,51,52,53]. They can always be reduced to *affine diagrams* (i.e., hyperplane bisectors) which can be computed either as equivalent *power diagrams* or by generalizing the Euclidean paraboloid lifting procedure by choosing the potential function $(x, F(x))$ instead of the paraboloid [50]. Statistical Voronoi diagrams can also be used for *multiple* class hypothesis testing: Figure 2 illustrates a geometric characterization of the Chernoff distance of a set of n distributions belonging to the same exponential families. Refer to [54] for further explanations.

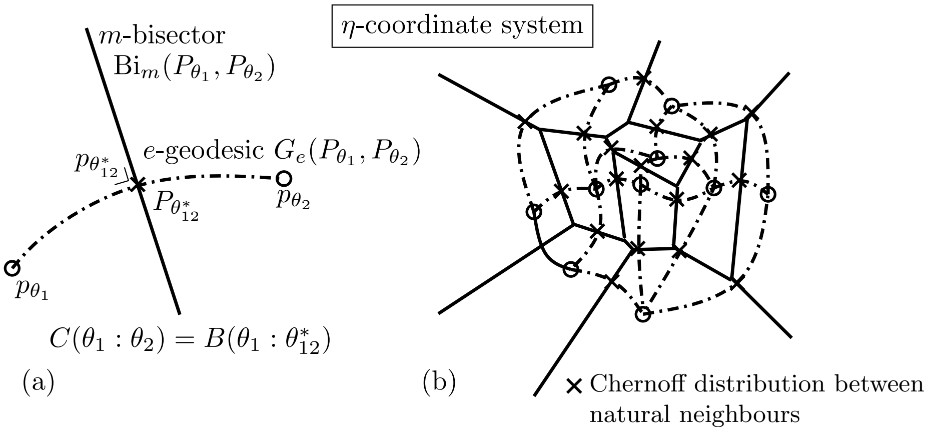
m-bisector
$\mathrm{Bi}_m(P_{\theta_1}, P_{\theta_2})$
e-geodesic $G_e(P_{\theta_1}, P_{\theta_2})$
$p_{\theta^*_{12}}$
$P_{\theta^*_{12}}$
p_{θ_2}
p_{θ_1}
$C(\theta_1 : \theta_2) = B(\theta_1 : \theta^*_{12})$
(a)
η-coordinate system
(b)
× Chernoff distribution between natural neighbours

Fig. 2. Geometry of the best error exponent in Bayesian classification [54]. Binary hypothesis (a): The Chernoff distance is equal to the Kullback-Leibler divergence from the midpoint distribution $P_{\theta^*_{12}}$ to the extremities, where the midpoint distribution $P_{\theta^*_{12}}$ ($\times$) is obtained as the left-sided KL projection of the sites to their bisector [55]. (b) Multiple hypothesis testing: The Chernoff distance is the minimum of pairwise Chernoff distance that can be deduced from statistical Voronoi diagram by inspecting all Chernoff distributions ($\times$) lying on $(d-1)$-faces. Both drawings illustrated in the η-coordinate system where m-bisectors are hyperplanes.

6 Conclusion and perspectives

We concisely reviewed the principles of computational information geometry for pattern learning and recognition on statistical manifolds: We consider statistical patterns whose distributions are either represented by atomic distributions (parametric models, say, of an exponential family), mixtures thereof (semi-parametric models), or kernel density estimations (non-parametric models). Those statistical pattern representations need to be estimated from datasets. We presented a geometric framework to learn and process those statistical patterns by embedding them on statistical manifolds. A statistical pattern is then represented either by a single point (parametric model), a k-weighted point set or a n-point set on the statistical manifold. To discriminate between patterns, we introduced the notion of statistical distances, and presented a genesis that yielded the family of α-divergences. We described the two notions of statistical invariances on statistical manifolds: invariance by sufficient statistic and invariance by 1-to-1 reparameterization of distribution parameters. We then introduced two kinds of statistical manifolds that fulfills the statistical invariance: The Rao manifolds based on Riemannian geometry using the Fisher information matrix as the underlying metric tensor, and the Amari-Chentsov dually flat manifolds based on the convex duality induced by a convex functional generator. We then explained why the usual lack of closed-form geodesic expression for Rao manifolds yields a limited algorithmic toolbox. By contrast, the explicit dual geodesics of Amari-Chentsov manifolds provides a handy framework to extend the Euclidean algorithmic toolbox. We illustrated those concepts by reviewing the Voronoi diagrams (and dual Delaunay triangulations), and considered simplifying mixtures or KDEs using clustering techniques. In particular, in the Amari-Chentsov manifolds, we can compute using either the primal, dual, or mixed coordinate systems. This offers many strategies for efficient computing. For the exponential family manifolds, we explained the bijection between exponential families, dual Bregman divergences and quasi-arithmetic means [10].

We would like to conclude with perspectives for further work. To begin with, let us say that there are several advantages to think "geometrically":

- First, it allows to use simple concepts like line segments, balls, projections to describe properties or algorithms. The language of geometry gives special affordances for human thinking. For example, to simplify a mixture of exponential families to a single component amount to project the mixture model onto the exponential family manifold (depicted in Figure 1). Algorithmically, this projection is performed by computing a barycenter.
- Second, sometimes we do not have analytical solution but nevertheless we can still describe geometrically exactly where the solution is. For example, consider the Chernoff information of two distributions: It is computed as the Kullback-Leibler divergence from the mid-distribution to the extremities (depicted in Figure 2). The mid-distribution is the unique distribution that is at the intersection of the exponential geodesic with the mixture bisector.

We implemented those various algorithms in the jMEF[4] [56] or PyMEF[5] [57] software libraries.

To quote mathematician Jules H. Poincaré: "One geometry cannot be more true than another; it can only be more convenient". We have exemplified this quote by showing that geometry is not absolute nor ultimate: Indeed, we have shown two kinds of geometries for handling statistical manifolds: Rao Riemannian manifolds and Amari-Chentsov dual affine manifolds. We also presented several *mathematical tricks* that yielded computational convenience: Bounding the intersection similarity measure with quasi-arithmetic means extends the α-divergences. Besides the Rao and Amari-Chentsov manifolds, we can also consider Finsler geometry [58] or Hilbert spherical geometry in infinite dimensional spaces to perform statistical pattern recognition. Non-extensive entropy pioneered by Tsallis also gave birth to *deformed exponential families* that have been studied using conformal geometry. See also the infinite-dimensional exponential families and Orlicz spaces [59], the optimal transport geometry [60], the symplectic geometry, Kähler manifolds and Siegel domains [61], the Geometry of proper scoring rules [62], the quantum information geometry [63], etc, etc. This raises the question of knowing which geometry to choose? For a specific application, we can study and compare experimentally say Rao vs. Amari-Chentsov manifolds. However, we need deeper axiomatic understandings in future work to (partially) answer this question. For now, we may use Rao manifolds if we require metric properties of the underlying distance, or if we want to use the triangular inequality to improve k-means clustering or nearest-neighbor searches. Some applications require to consider symmetric divergences: We proposed a parametric family of symmetric divergences [64] including both the Jeffreys divergence and the Jensen-Shannon divergence, and described the centroid computations with respect to that class of distances.

Geometry offers many more possibilities to explore in the era of big data analytics as we are blinded with numbers and need to find rather qualitative invariance of the underlying space of data. There are many types of geometries to explore or invent as mothers of models. Last but not least, we should keep in mind statistician George E. P. Box quote: "Essentially, all models are wrong, but some are useful." When it comes to data spaces, we also believe that all geometries are wrong, but some are useful.

References

1. Jain, Anil K. and Duin, Robert P. W. and Mao, Jianchang: Statistical pattern recognition: A review. IEEE Trans. Pattern Anal. Mach. Intell. **22** 4–37
2. Harold Cramér: Mathematical Methods of Statistics. Princeton Landmarks in mathematics (1946)
3. Maurice Fréchet: Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. Review of the International Statistical Institute **11** 182–205 published in IHP Lecture in 1939.

[4] `http://www.lix.polytechnique.fr/~nielsen/MEF/`

[5] `http://www.lix.polytechnique.fr/~schwander/pyMEF/`

4. Calyampudi Radhakrishna Rao: Information and the accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society **37** 81–89
5. Frank Nielsen: In . : Connected at Infinity II: A selection of mathematics by Indians. Cramér-Rao lower bound and information geometry (Hindustan Book Agency (Texts and Readings in Mathematics, TRIM)) arxiv 1301.3578.
6. Arthur Pentland Dempster and Nan M. Laird and Donald B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) **39** 1–38
7. Keinosuke Fukunaga: Introduction to statistical pattern recognition. Academic Press Professional, Inc. (1990) 2nd ed. (1st ed. 1972).
8. Paolo Piro and Frank Nielsen and Michel Barlaud: Tailored Bregman ball trees for effective nearest neighbors. In: European Workshop on Computational Geometry (EuroCG), LORIA, Nancy, France, IEEE (2009)
9. Frank Nielsen and Paolo Piro and Michel Barlaud: Bregman vantage point trees for efficient nearest neighbor queries. In: Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME). 878–881
10. Richard Nock and Frank Nielsen: Fitting the smallest enclosing bregman balls. In: 16th European Conference on Machine Learning (ECML). 649–656
11. Nielsen, Frank and Nock, Richard: On the smallest enclosing information disk. Inf. Process. Lett. **105** 93–97
12. Nielsen, Frank and Nock, Richard: On approximating the smallest enclosing Bregman balls. In: ACM Symposium on Computational Geometry (SoCG), ACM Press (2006)
13. Marc Arnaudon and Frank Nielsen: On approximating the Riemannian 1-center. Computational Geometry **46** 93 – 104
14. Frank Nielsen and Richard Nock: Approximating smallest enclosing balls with applications to machine learning. Int. J. Comput. Geometry Appl. **19** 389–414
15. Syed Mumtaz Ali and Samuel David Silvey: A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society, Series B **28** 131–142
16. Imre Csiszár: Information-type measures of difference of probability distributions and indirect observation. Studia Scientiarum Mathematicarum Hungarica **2** 229–318
17. Cover, Thomas M. and Thomas, Joy A.: Elements of information theory. Wiley-Interscience, New York, NY, USA (1991)
18. Frank Nielsen: Closed-form information-theoretic divergences for statistical mixtures. In: International Conference on Pattern Recognition (ICPR). (2012)
19. Wu, Jianxin and Rehg, James M.: Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: ICCV. (2009)
20. Frank Nielsen and Vincent Garcia: Statistical exponential families: A digest with flash cards (2009) arXiv.org:0911.4863.
21. Martin E. Hellman and Josef Raviv: Probability of error, equivocation and the Chernoff bound. IEEE Transactions on Information Theory **16** 368–372
22. Frank Nielsen and Sylvain Boltz: The Burbea-Rao and Bhattacharyya centroids. IEEE Transactions on Information Theory **57** 5455–5466
23. Shun-ichi Amari and Hiroshi Nagaoka: Methods of Information Geometry. Oxford University Press (2000)
24. Qiao, Yu and Minematsu, Nobuaki: A study on invariance of f-divergence and its application to speech recognition. Transactions on Signal Processing **58** 3884–3890

25. María del Carmen Pardo and Igor Vajda: About distances of discrete distributions satisfying the data processing theorem of information theory. IEEE Transactions on Information Theory **43** 1288–1293
26. Shun-ichi Amari: alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes. IEEE Transactions on Information Theory **55** 4925–4931
27. Elena Alexandra Morozova and Nikolai Nikolaevich Chentsov: Markov invariant geometry on manifolds of states. Journal of Mathematical Sciences **56** 2648–2669
28. Ronald Aylmer Fisher: On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London, A **222** 309–368
29. Nikolai Nikolaevich Chentsov: Statistical Decision Rules and Optimal Inferences. Transactions of Mathematics Monograph, numero 53 (1982) Published in russian in 1972.
30. Adrian Peter and Anand Rangarajan: A new closed-form information metric for shape analysis. Volume. 1. 249–256
31. Colin Atkinson and Ann F. S. Mitchell: Rao's distance measure. Sankhya A **43** 345–365
32. Miroslav Lovric and Maung Min-Oo and Ernst A Ruh: Multivariate normal distributions parametrized as a Riemannian symmetric space. Journal of Multivariate Analysis **74** 36 – 48
33. Olivier Schwander and Frank Nielsen: Model centroids for the simplification of kernel density estimators. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 737–740
34. Marc Arnaudon and Frank Nielsen: Medians and means in Finsler geometry. CoRR **abs/1011.6076** (2010)
35. Frank Nielsen and Richard Nock: Hyperbolic Voronoi diagrams made easy. Volume. 1., Los Alamitos, CA, USA, IEEE Computer Society 74–80
36. Frank Nielsen and Richard Nock: The hyperbolic voronoi diagram in arbitrary dimension. CoRR **abs/1210.8234** (2012)
37. Xavier Pennec: In F. Nielsen (Ed). Emerging Trends in Visual Computing (ETVC). 347–386
38. Banerjee, Arindam and Merugu, Srujana and Dhillon, Inderjit S. and Ghosh, Joydeep: Clustering with Bregman divergences. Journal of Machine Learning Research **6** 1705–1749
39. Barndorff-Nielsen, O.E.: Information and exponential families: in statistical theory. Wiley series in probability and mathematical statistics: Tracts on probability and statistics. Wiley (1978)
40. Bogdan, Krzysztof and Bogdan, Małgorzata: On existence of maximum likelihood estimators in exponential families. Statistics **34** 137–149
41. Frank Nielsen: k-MLE: A fast algorithm for learning statistical mixture models. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE (2012) preliminary, technical report on arXiv.
42. Olivier Schwander and Frank Nielsen and Aurélien Schutz and Yannick Berthoumieu: k-MLE for mixtures of generalized Gaussians. In: International Conference on Pattern Recognition (ICPR). (2012)
43. Olivier Schwander and Frank Nielsen: Fast learning of Gamma mixture models with k-MLE. In: International Workshop on Similarity-Based Pattern Analysis and Recognition (SIMBAD). (2013) this proceedings.
44. Christophe Saint-Jean and Frank Nielsen: A new implementation of k-MLE for mixture modelling of Wishart distributions. In: Geometric Sciences of Information (GSI). (2013)

45. Olivier Schwander and Frank Nielsen: In Bhatia and Nielsen (Eds). Learning Mixtures by Simplifying Kernel Density Estimators, Matrix Information Geometry. 403–426
46. Frank Nielsen and Richard Nock: Sided and symmetrized Bregman centroids. IEEE Transactions on Information Theory **55** 2882–2904
47. Vincent Garcia and Frank Nielsen and Richard Nock: Levels of details for Gaussian mixture models. Volume. 2. 514–525
48. Baba Vemuri and Meizhu Liu and Shun-ichi Amari and Frank Nielsen: Total Bregman divergence and its applications to DTI analysis. IEEE Transactions on Medical Imaging (2011) 10.1109/TMI.2010.2086464.
49. Meizhu Liu and Baba C. Vemuri and Shun-ichi Amari and Frank Nielsen: Shape retrieval using hierarchical total Bregman soft clustering. Transactions on Pattern Analysis and Machine Intelligence (2012)
50. Boissonnat, Jean-Daniel and Nielsen, Frank and Nock, Richard: Bregman Voronoi diagrams. Discrete Comput. Geom. **44** 281–307
51. Nielsen, Frank and Boissonnat, Jean-Daniel and Nock, Richard: On Bregman Voronoi diagrams. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. SODA '07, Philadelphia, PA, USA, Society for Industrial and Applied Mathematics 746–755
52. Nielsen, Frank and Boissonnat, Jean-Daniel and Nock, Richard: Visualizing Bregman Voronoi diagrams. In: Proceedings of the twenty-third annual symposium on Computational geometry. SCG '07, New York, NY, USA, ACM 121–122
53. Frank Nielsen and Richard Nock: Jensen-Bregman Voronoi diagrams and centroidal tessellations. In: International Symposium on Voronoi Diagrams (ISVD). 56–65
54. Frank Nielsen: Hypothesis testing, information divergence and computational geometry. Geometric Sciences of Information (GSI) (2013)
55. Frank Nielsen: An information-geometric characterization of Chernoff information. IEEE Signal Processing Letters (SPL) **20** 269–272
56. Vincent Garcia and Frank Nielsen: Simplification and hierarchical representations of mixtures of exponential families. Signal Processing (Elsevier) **90** 3197–3212
57. Olivier Schwander and Frank Nielsen: PyMEF - A framework for exponential families in Python. In: IEEE/SP Workshop on Statistical Signal Processing (SSP). (2011)
58. Zhongmin Shen: Riemann-Finsler geometry with applications to information geometry. Chinese Annals of Mathematics **27B** 73–94
59. Alberto Cena and Giovanni Pistone: Exponential statistical manifold. Annals of the Institute of Statistical Mathematics **59** 27–56
60. Gangbo, Wilfrid and McCann, Robert J.: The geometry of optimal transportation. Acta Math. **177** 113–161
61. Barbaresco, Frédéric: Interactions between Symmetric Cone and Information Geometries: Bruhat-Tits and Siegel Spaces Models for High Resolution Autoregressive Doppler Imagery. In Emerging Trends in Visual Computing. Volume. 5416 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg 124–163
62. Alexander Philip Dawid: The geometry of proper scoring rules. Annals of the Institute of Statistical Mathematics **59** 77–93
63. Matheus R. Grasselli and Raymond F. Streater: On the uniqueness of the Chentsov metric in quantum information geometry. Infinite Dimensional Analysis, Quantum Probability and Related Topics **4** 173–181 arXiv.org:math-ph/0006030.
64. Frank Nielsen: A family of statistical symmetric divergences based on Jensen's inequality. CoRR **abs/1009.4004** (2010)