Distances, divergences, statistical divergences and diversities

Frank Nielsen
Sony Computer Science Laboratories Inc.
Tokyo, Japan

August 2021
(Last updated November 12, 2021)

This document is also available in PDF: `Distance.pdf`

This is a *working document* which will be (hopefully frequently) updated with materials concerning the discrepancies between two distributions/parameters or the diversities of a set of distributions/parameters. There are many synonyms in the literature to measure the difference between two objects: Metrics, Distances, Discrepancies, deviations, deviances, dissimilarities, divergences, contrast functions or yokes (on product manifolds), etc. Diversities generalize 2-point distances by measuring the dispersion of a set of n objects, usually using a centrality notion.

In mathematics, a distance is often considered to be a metric distance in a metric space which satisfies the following properties:

There is confusion in the literature where distance is also used as a synonym of a dissimilarity measure.

In information theory and statistics, we measure deviations between a probability measure and another probability measure using a statistical divergence.

In information geometry, a divergence is a smooth dissimilarity measure which shall satisfy the following conditions:

Divergences were formerly called contrast functions, a dualistic structure of information geometry can be built from a divergence.

Contents

[width=0.75]FigIpe-skewJS.pdf

Figure 1: Skewed Jensen divergences visualized as vertical convexity gaps.

1 Jensen divergences and Bregman divergences

1.1 Skewed Jensen and Bregman divergences

$$\lim_{\alpha\rightarrow 0} \mathrm{sJ}_{F,\alpha}(\theta_1:\theta_2) = B_F(\theta_1:\theta_2)$$

$$\lim_{\alpha\rightarrow 1} \mathrm{sJ}_{F,\alpha}(\theta_1:\theta_2) = B_F(\theta_2:\theta_1)$$

If $\alpha<0$ or $\alpha>1$, we can measure the gap $F((1-\alpha)\theta_1+\alpha\theta_2)-((1-\alpha)F(\theta_1)+\alpha F(\theta_2)) = -J_F(\theta_1,\theta_2)\geq 0$. See Figure 1 Thus we can define the scaled α-skew Jensen divergence for $\alpha\in\mathbb{R}\backslash\{0,1\}$ as:

$$_F(\theta_1:\theta_2) = \frac{1}{\alpha(1-\alpha)} J_F(\theta_1:\theta_2)\geq 0.$$

1.2 Relationships with statistical distances between densities of an exponential family

1.3 Generalizations of Bregman and Jensen divergences

2 Invariant f-divergences

A f-divergence [26, 7, 1, 8] $I_f[p:q]$ is a dissimilarity measure between probability distributions defined for a convex generator $f(u)$:

$$I_f[p:q] = \int p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}\mu(x).$$

Using Jensen's inequality, we have $I_f[p:q]\geq f(1)$. Thus we ask for generators satisfying $f(1)=0$. Moreover, in order to have $I_f[p:q]=0$ iff. $p=q$ (μ-almost everywhere), we require $f(u)$ to be strictly convex at 1. The f-divergences include many well-known statistical distances listed with their generators:

f-divergence	Formula $I_f[p:q]$	generator $f(u)$
Total variation (metric)	$\frac{1}{2}\int \lvert p(x)-q(x)\rvert \mathrm{d}\mu(x)$	$\frac{1}{2}\lvert u-1\rvert$
Squared Hellinger	$\int(\sqrt{p(x)}-\sqrt{q(x)})^2\mathrm{d}\mu(x)$	$(\sqrt{u}-1)^2$
Pearson χ^2_P	$\int \frac{(q(x)-p(x))^2}{p(x)}\mathrm{d}\mu(x)$	$(u-1)^2$
Neyman χ^2_N	$\int \frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}\mu(x)$	$\frac{(1-u)^2}{u}$
Kullback-Leibler	$\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$	$-\log u$
reverse Kullback-Leibler	$\int q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\mu(x)$	$u\log u$
Jeffreys divergence	$\int(p(x)-q(x))\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$	$(u-1)\log u$
α-divergence	$\frac{4}{1-\alpha^2}(1-\int p^{\frac{1-\alpha}{2}}(x)q^{1+\alpha}(x)\mathrm{d}\mu(x))$	$\frac{4}{1-\alpha^2}(1-u^{\frac{1+\alpha}{2}})$
Jensen-Shannon	$\frac{1}{2}\int(p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)})\mathrm{d}\mu(x)$	$-(u+1)\log\frac{1+u}{2}+u\log u$

Two f-divergences I_{f_1} and I_{f_2} are equivalent iff. $f_1(u)=f_2(u)+\lambda(u-1)$ for any $\lambda\in\mathbb{R}$. A symmetric f-divergence is bounded (e.g., the Jensen-Shannon divergence or the total variation) iff. $f(0)<\infty$. The Jeffreys divergence is an unbounded f-divergence. The dual f-divergence $I_f{}^*[p:q]:=I_f[p:q]$ is a f-divergence for the dual generator $f^*(u)=uf\left(\frac{1}{u}\right)$ (or conjugate generator). Thus symmetric f-divergences (e.g., the Jeffreys divergence, Hellinger divergence, or the Jensen-Shannon divergence) satisfies the functional equality $f(u)=uf(1/u)$. The f-divergences are joint convex and satisfies the information monotonicity property: $I_f[p_{|\mathcal{Y}}:q_{|\mathcal{Y}}]\leq I_f[p:q]$ for any partition $\mathcal{Y}$ of $\mathcal{X}$ (see lumping [9]). A statistical divergence is said separable iff. it can be rewritten as $D[p:q]=\int D_1(p(x):q(x))\mathrm{d}\mu(x)$, where D_1 is a scalar divergence. The f-divergences are the only divergences which are separable and satisfies the information information monotonicity [3] (except the "curious case" [17] of binary alphabets $\mathcal{X}$). A f-divergence is said standard [3] when $f''(1)=1$. The local Taylor expansion [6] of $I_f[p_{\theta_1}:p_{\theta_2}]$-divergences between two parametric divergences is related to the Fisher information matrix $I(\theta)=E_{p_\theta}\left[\nabla\log p_\theta(x)(\nabla\log p_\theta(x))^\top\right]$ as follows:

$$I_f[p_{\theta_1}:p_{\theta_2}]=\frac{1}{2}(\theta_2-\theta_1)^\top I_\theta(\theta_1)(\theta_2-\theta_1)+o(\|(\theta_2-\theta_1\|^2).$$

Thus we have $I_f[p_\theta:p_{\theta+\mathrm{d}\theta}]=\frac{1}{2}\mathrm{d}\theta^\top I(\theta)\mathrm{d}\theta$ for a standard f-divergence (with $f''(1)=1$). The following metric distance D^Q is called the Mahalanobis distance [24][1]:

$$D^{Q(\theta_1,\theta_2)=\sqrt{(\theta_2-\theta_1)^\top Q(\theta_2-\theta_1)}.}$$

The Mahalanobis distance generalizes the Euclidean distance (expressed in the Cartesian coordinate system) using $Q=I$, the identity matrix. When $\Sigma=(\sigma_{11}^2,\ldots,\sigma_{DD}^2)$ with $\sigma_{ii}=\sigma_i^2$, we have

$$D^{\Sigma^{-1}(\theta_1,\theta_2)=\sum_{i=1}^D\frac{(\theta_2^i-\theta_1^i)^2}{\sigma_i^2}.}$$

Thus we have $I_f[p_{\theta_1}:p_{\theta_2}]=\frac{1}{2}D^{I_\theta(\theta_1)(\theta_1,\theta_2)^2+o(\|(\theta_2-\theta_1\|^2)}$. For $\theta_1=\theta$ and $\theta_2=\theta+\mathrm{d}\theta$, the half squared Mahalanobis distance can be interpreted as the squared Riemannian infinitesimal length element: $I_f[p_\theta:p_{\theta+\mathrm{d}\theta}]=\frac{1}{2}\mathrm{d}\theta^\top I_\theta(\theta)\mathrm{d}\theta=^2_\theta$.

[1]Mahalanobis defined that distance for $Q=\Sigma^{-1}\succ 0$, the inverse of a covariance matrix.

A f-divergence between any two densities p_{θ_1} and p_{θ_2} can be expressed as a Taylor series [40] when $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} < 1 + r_f$ where r_f is the convergence radius of the analytic generator $f \in C^\omega$ and $\frac{p_{\theta_1}}{p_{\theta_2}} \leq C$:

$$I_f[p_{\theta_1} : p_{\theta_2}] = \sum_{n=2}^{\infty} a_n \int_X \left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} - 1\right)^n p_{\theta_1}(x) \mathrm{d}\mu(x).$$

Otherwise, the Taylor series diverge.

By introducing the higher-order chi divergences [37]

$$D_{\chi,n}[p : q] = \int \frac{(p(x) - q(x))^n}{(q(x))^{n-1}(x)} \mathrm{d}\mu(x)$$

which are proper divergences for even integers and only pseudo-distances for odd orders, we rewrite the Taylor series of f-divergences as:

$$I_f[p_{\theta_1} : p_{\theta_2}] = \sum_{n=2}^{\infty} a_n D_{\chi,n}[p_{\theta_1} : p_{\theta_2}].$$

The higher-order chi divergences [37] between densities of an exponential family are available in closed-form provided that the natural parameter space is affine (e.g., isotropic Gaussian family or Poisson family).

To illustrate the Taylor series, consider the Poisson family, and let us express the Taylor series for the Jensen-Shannon divergence (generator $f_{\mathrm{JS}}(u) = -(u+1)\log\frac{1+u}{2} + u \log u$) between two Poisson distributions with parameters $\lambda_1 = 0.6$ and $\lambda_2 = 0.3$. We have the higher-order chi divergences between Poisson distributions expressed in closed-form as follows [37]:

$$D_{\chi,k}[p_{\lambda_1} : p_{\lambda_2}] = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} e^{\lambda_1^{1-j}\lambda_2^{j} - ((1-j)\lambda_1 + j\lambda_2)}$$

Furthermore, $\frac{p_{\lambda_1}(x)}{p_{\lambda_2}(x)} = \frac{\lambda_1^k e^{-\lambda_1}}{\lambda_2^k e^{-\lambda_2}} < C$

We have

$$I_f[p : q] \sim \frac{f''(1)}{2} D_{\chi^2_N}[p : q],$$

where $D_{\chi^2}[p : q] = \int \frac{(p(x)-q(x))^2}{q(x)} \mathrm{d}\mu(x)$ is the chi-squared divergence. Therefore, we have

$$D_{\mathrm{KL}}[p : q] \sim \frac{1}{2} D_{\chi^2}[p : q]. \tag{1}$$

On the finite-dimensional probability simplex, the Kullback-Leibler divergence is the only statistical divergence which belongs to both the f-divergences and the Bregman divergences [2]. When considering the f-divergences to positive measures, the intersection of the f-divergences and the Bregman divergences are the α-divergences [2].

For the parametric family of Cauchy distributions, the f-divergences are always symmetric and can be expressed as a function of the chi-squared divergence [40]:

$$I_f[p^{\mathrm{Cauchy}}_{l_1,s_1} : p^{\mathrm{Cauchy}}_{l_2,s_2}] = I_f[p^{\mathrm{Cauchy}}_{l_2,s_2} : p^{\mathrm{Cauchy}}_{l_1,s_1}] = h_f(D_{\chi^2}[p^{\mathrm{Cauchy}}_{l_1,s_1} : p^{\mathrm{Cauchy}}_{l_2,s_2}]),$$

where

$$D_{\chi^2}[p^{\mathrm{Cauchy}}_{l_1,s_1} : p^{\mathrm{Cauchy}}_{l_2,s_2}] = \frac{(l_2-l_1)^2+(s_2-s_1)^2}{2s_1s_2}$$

and

$$p^{\mathrm{Cauchy}}_{l,s}(x) := \frac{1}{\pi s\left(1+\left(\frac{x-l}{s}\right)^2\right)} = \frac{s}{\pi(s^2+(x-l)^2)}.$$

3 Distances and means

4 Statistical distances between densities with computationally intractable normalizers

Consider a density $p(x) = \frac{\tilde{p}(x)}{Z_p}$ where $\tilde{p}(x)$ is an unnormalized *computable* density and $Z_p = \int p(x)\mathrm{d}\mu(x)$ the *computationally intractable* normalizer (also called in statistical physics the partition function or free energy). A statistical distance $D[p_1 : p_2]$ between two densities $p_1(x) = \frac{\tilde{p}_1(x)}{Z_{p_1}}$ and $p_2(x) = \frac{\tilde{p}_2(x)}{Z_{p_2}}$ with computationally intractable normalizers Z_{p_1} and Z_{p_2} is said *projective* (or two-sided *homogeneous*) if and only if

$$\forall \lambda_1 > 0, \lambda_2 > 0, \quad D[p_1 : p_2] = D[\lambda_1 p_1 : \lambda_2 p_2].$$

In particular, letting $\lambda_1 = Z_{p_1}$ and $\lambda_2 = Z_{p_2}$, we have

$$D[p_1 : p_2] = D[\tilde{p}_1 : \tilde{p}_2].$$

Notice that the rhs. does not rely on the computationally intractable normalizers. These projective distances are useful in statistical inference based on minimum distance estimators [5] (see next Section).

Here are a few statistical projective distances:

- **γ-divergences** ($\gamma > 0$) [18, 13]:

  $$D_\gamma[p : q] := \log\left(\int_{\mathbb{R}} q^{\alpha+1}\right) - \left(1+\frac{1}{\alpha}\right)\log\left(\int_{\mathbb{R}} q^\alpha p\right) + \frac{1}{\alpha}\log\left(\int_{\mathbb{R}} p^{\alpha+1}\right), \quad \gamma \geq 0$$

  When $\gamma \to 0$, we have [13] $D_\gamma[p : q] = D_{\mathrm{KL}}[p : q]$, the Kullback-Leibler divergence (KLD). For example, we can estimate the KLD between two densities of an exponential-polynomial family by Monte Carlo stochastic integration of the γ-divergence for a small value of γ [38].

  The γ-divergences (projective, Bregman-type=Cross-entropy-entropy) and the density power divergence [4] (non-projective, Bregman-type divergence):

  $$D^{\mathrm{dpd}}_\alpha[p : q] := \int_{\mathbb{R}} q^{\alpha+1} - \left(1+\frac{1}{\alpha}\right)\int_{\mathbb{R}} q^\alpha p + \frac{1}{\alpha}\int_{\mathbb{R}} p^{\alpha+1}, \quad \alpha \geq 0,$$

  can be encapsulated into the family of Φ-power divergences [49] (functional density power divergence class):

  $$D_{\phi,\alpha}[p : q] := \phi\left(\int_{\mathbb{R}} q^{\alpha+1}\right) - \left(1+\frac{1}{\alpha}\right)\phi\left(\int_{\mathbb{R}} q^\alpha p\right) + \frac{1}{\alpha}\phi\left(\int_{\mathbb{R}} p^{\alpha+1}\right), \quad \alpha \geq 0,$$

where $\phi(e^x)$ convex and strictly increasing, ϕ continuous and twice continously differentiable with finite second order derivatives. We have $D_{\phi,0}[p:q] = \phi'(1)\int_{\mathbb{R}} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x) = \phi'(1)D_{\mathrm{KL}}[p:q]$.

- **Cauchy-Schwarz divergence** [16] (CSD, projective)

$$D_{\mathrm{CS}}[p:q] = -\log\left(\frac{\int p(x)q(x)\mathrm{d}\mu(x)}{\sqrt{\int p(x)^2\mathrm{d}\mu(x)\int q(x)^2\mathrm{d}\mu(x)}}\right) = D_{\mathrm{CS}}[\lambda_1 p:\lambda_2 q], \forall\lambda_1>0,\lambda_2>0,$$

and **Hölder divergences** [46] (HD, projective, which generalizes the CSD):

$$D^{\text{Hölder}}_{\alpha,\gamma}[p:q] = -\log\left(\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha}q(x)^{\gamma/\beta}\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\gamma}\mathrm{d}x\right)^{1/\alpha}\left(\int_{\mathcal{X}} q(x)^{\gamma}\mathrm{d}x\right)^{1/\beta}}\right), \quad \frac{1}{\alpha}+\frac{1}{\beta}=1.$$

We have

$$\forall\lambda_1>0,\lambda_2>0, D^{\text{Hölder}}_{\alpha,\gamma}[\lambda_1 p:\lambda_2 q] = D^{\text{Hölder}}_{\alpha,\gamma}[p:q],$$

and

$$D^{\text{Hölder}}_{2,2}[p:q] = D_{\mathrm{CS}}[p:q].$$

Hölder divergences between two densities p_{θ_p} and p_{θ_q} of an exponential family with cumulant function $F(\theta)$ is available in closed-form [46]:

$$D^{\text{Hölder}}_{\alpha,\gamma}[p:q] = \frac{1}{\alpha}F(\gamma\theta_p) + \frac{1}{\beta}F(\gamma\theta_q) - F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right)$$

The CSD is available in closed-form between mixtures of an exponential family with a conic natural parameter [28]: This includes the case of Gaussian mixture models [19].

- **Hilbert distance** [45] (projective): Consider two probability mass functions $p=(p_1,\ldots,p_d)$ and $q=(q_1,\ldots,q_d)$ of the d-dimensional probability simplex. Then the Hilbert distance is

$$D^{\mathrm{Hilbert}}[p:q] = \log\left(\frac{\max_{i\in\{1,\ldots,d\}}\frac{p_i}{q_i}}{\min_{j\in\{1,\ldots,d\}}\frac{p_j}{q_j}}\right).$$

We have

$$\forall\lambda_1>0,\lambda_2>0, D^{\mathrm{Hilbert}}[\lambda_1 p:\lambda_2 q] = D^{\mathrm{Hilbert}}[p:q].$$

The Hilbert projective simplex distance can be extended to the cone of positive-definite matrices [45] (and its subspace of correlation matrices called the elliptope) as follows:

$$D^{\mathrm{Hilbert}}[P:Q] = \log\left(\frac{\lambda_{\max}(PQ^{-1})}{\lambda_{\min}(PQ^{-1})}\right),$$

where $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ denote the largest and smallest eigenvalue of matrix X, respectively.

5 Statistical distances between empirical distributions and densities with computationally intractable normalizers

When estimating the parameter $\hat{\theta}$ for a parametric family of distributions $\{p_\theta\}$ from i.i.d. observations $\mathcal{S} = \{x_1, \ldots, x_n\}$, we can define a minimum distance estimator (MDE):

$$\hat{\theta} = \arg\min_\theta D[p_{\mathcal{S}} : p_\theta],$$

where $p_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical distribution (normalized). Thus we need only a right-sided projective divergence to estimate models with computationally intractable normalizers. For example, the Maximum Likelihood Estimator (MLE) is a MDE wrt. the KLD:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_\theta D_{\mathrm{KL}}[p_{\mathcal{S}} : p_\theta].$$

It is thus interesting to study the impact of the choice of the distance D to the properties of the corresponding estimator (e.g., γ-divergences yields provably robust estimators [13]).

- **Hyvärinen divergence** [14] (also called **Fisher divergence** or Fisher relative information [47]):

  $$D^{\text{Hyvärinen}}[p : p_\theta] := \frac{1}{2}\int \|\nabla_x \log p(x) - \nabla_x \log p_\theta(x)\|^2\, p(x)\mathrm{d}x.$$

  The Hyvarinen divergence has been extended for order-α Hyvarinen divergences [32] (for $\alpha > 0$):

  $$D_\alpha^{\text{Hyvärinen}}[p : q] := \frac{1}{2}\int p(x)^\alpha \left(\nabla_x \log p(x) - \nabla_x \log q(x)\right)^2 \mathrm{d}x, \quad \alpha > 0.$$

  The Fisher divergence is related to the Kullback-Leibler divergence [53] as follows:

  $$D_{\mathrm{KL}}[p : q] = \int_0^\infty D_{\mathrm{Fisher}}[p * (0, \lambda I) : q * (0, \lambda I)]\,,$$

  where $(f * g)(x) = \int f(y)g(x-y)$ denotes the convolution of densties. Thus convergence wrt Fisher divergence is stronger than convergence wrt KLD.

6 The Jensen-Shannon divergence and some generalizations

6.1 Origins of the Jensen-Shannon divergence

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measure space, and $(w_1, P_1), \ldots, (w_n, P_n)$ be n weighted probability measures dominated by a measure μ (with $w_i > 0$ and $\sum w_i = 1$). Denote by $\mathcal{P} := \{(w_1, p_1), \ldots, (w_n, p_n)\}$ the set of their weighted Radon-Nikodym densities $p_i = \frac{\mathrm{d}P_i}{\mathrm{d}\mu}$ with respect to μ.

A *statistical divergence* $D[p : q]$ is a measure of dissimilarity between two densities p and q (i.e., a 2-point distance) such that $D[p : q] \geq 0$ with equality if and only if $p(x) = q(x)$ μ-almost everywhere. A *statistical diversity index* $D(\mathcal{P})$ is a measure of variation of the weighted densities in

$\mathcal{P}$ related to a measure of centrality, i.e., a n-point distance which generalizes the notion of 2-point distance when $\mathcal{P}_2(p,q) := \{(\frac{1}{2}, p_1), (\frac{1}{2}, p_2)\}$:

$$D[p:q] := D(\mathcal{P}_2(p,q)).$$

The fundamental measure of dissimilarity in information theory is the *I-divergence* (also called the *Kullback-Leibler divergence*, KLD, see Equation (2.5) page 5 of [20]):

$$D_{\mathrm{KL}}[p:q] := \int_{\mathcal{X}} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x).$$

The KLD is asymmetric (hence the delimiter notation ":" instead of ',') but can be symmetrized by defining the Jeffreys *J-divergence* (Jeffreys divergence, denoted by I_2 in Equation (1) in 1946's paper [15]):

$$D_J[p,q] := D_{\mathrm{KL}}[p:q] + D_{\mathrm{KL}}[q:p] = \int_{\mathcal{X}} (p(x) - q(x)) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}\mu(x).$$

Although symmetric, any positive power of Jeffreys divergence fails to satisfy the triangle inequality: That is, D_J^α is never a metric distance for any $\alpha > 0$, and furthermore D_J^α cannot be upper bounded.

In 1991, Lin proposed the asymmetric *K-divergence* (Equation (3.2) in [22]):

$$D_K[p:q] := D_{\mathrm{KL}}\left[p : \frac{p+q}{2}\right],$$

and defined the *L-divergence* by analogy to Jeffreys's symmetrization of the KLD (Equation (3.4) in [22]):

$$D_L[p,q] = D_K[p:q] + D_K[q:p].$$

By noticing that

$$D_L[p,q] = 2h\left[\frac{p+q}{2}\right] - (h[p] + h[q]),$$

where h denotes Shannon entropy (Equation (3.14) in [22]), Lin coined the (skewed) *Jensen-Shannon divergence* between two weighted densities $(1-\alpha, p)$ and (α, q) for $\alpha \in (0,1)$ as follows (Equation (4.1) in [22]):

$$D_{\mathrm{JS},\alpha}[p,q] = h[(1-\alpha)p + \alpha q] - (1-\alpha)h[p] - \alpha h[q]. \tag{2}$$

Finally, Lin defined the *generalized Jensen-Shannon divergence* (Equation (5.1) in [22]) for a finite weighted set of densities:

$$D_{\mathrm{JS}}[\mathcal{P}] = h\left[\sum_i w_i p_i\right] - \sum_i w_i h[p_i].$$

This generalized Jensen-Shannon divergence is nowadays called the *Jensen-Shannon diversity index*.

To contrast with the Jeffreys' divergence, the Jensen-Shannon divergence (JSD) $D_{\mathrm{JS}} := D_{\mathrm{JS},\frac{1}{2}}$ is upper bounded by $\log 2$ (does not require the densities to have the same support), and $\sqrt{D_{\mathrm{JS}}}$ is a metric distance [11, 12]. Lin cited precursor work [56, 23] yielding definition of the Jensen-Shannon divergence: The Jensen-Shannon divergence of Eq. 2 is the so-called "increments of entropy" defined in (19) and (20) of [56].

The Jensen-Shannon diversity index was also obtained very differently by Sibson in 1969 when he defined the *information radius* [52] of order α using Rényi α-means and Rényi α-entropies [50]. In particular, the information radius IR_1 of order 1 of a weighted set $\mathcal{P}$ of densities is a diversity index obtained by solving the following variational optimization problem:

$$\mathrm{IR}_1[\mathcal{P}] := \min_c \sum_{i=1}^n w_i D_{\mathrm{KL}}[p_i : c]. \tag{3}$$

Sibson solved a more general optimization problem, and obtained the following expression (term K_1 in Corollary 2.3 [52]):

$$\mathrm{IR}_1[\mathcal{P}] = h\left[\sum_i w_i p_i\right] - \sum_i w_i h[p_i] := D_{\mathrm{JS}}[\mathcal{P}].$$

Thus Eq. 3 is a variational definition of the Jensen-Shannon divergence.

6.2 Some extensions of the Jensen-Shannon divergence

- **Skewing the JSD.**

  The K-divergence of Lin can be skewed with a scalar parameter $\alpha \in (0,1)$ to give

  $$D_{K,\alpha}[p:q] := D_{\mathrm{KL}}\left[p : (1-\alpha)p + \alpha q\right]. \tag{4}$$

  Skewing parameter α was first studied in [21] (2001, see Table 2 of [21]). We proposed to unify the Jeffreys divergence with the Jensen-Shannon divergence as follows (Equation 19 in [27]):

  $$D^J_{K,\alpha}[p:q] := \frac{D_{K,\alpha}[p:q] + D_{K,\alpha}[q:p]}{2}. \tag{5}$$

  When $\alpha = \frac{1}{2}$, we have $D^J_{K,\frac{1}{2}} = D_{\mathrm{JS}}$, and when $\alpha = 1$, we get $D^J_{K,1} = \frac{1}{2} D_J$.

  Notice that

  $$D^{\alpha,\beta}_{\mathrm{JS}}[p;q] := (1-\beta) D_{\mathrm{KL}}[p : (1-\alpha)p + \alpha q] + \beta D_{\mathrm{KL}}[q : (1-\alpha)p + \alpha q]$$

  amounts to calculate

  $$h^\times[(1-\beta)p + \beta q : (1-\alpha)p + \alpha q] - ((1-\beta)h[p] + \beta h[q])$$

  where

  $$h^\times[p,q] := \int -p(x)\log q(x) \mathrm{d}\mu(x)$$

  denotes the *cross-entropy*. By choosing $\alpha = \beta$, we have $h^\times[(1-\beta)p + \beta q : (1-\alpha)p + \alpha q] = h[(1-\alpha)p + \alpha q]$, and thus recover the skewed Jensen-Shannon divergence of Eq. 2.

In [31] (2020), we considered a positive *skewing vector* $\alpha \in [0,1]^k$ and a unit positive weight w belonging to the standard simplex Δ_k, and defined the following *vector-skewed Jensen-Shannon divergence*:

$$D_{\mathrm{JS}}^{\alpha,w}[p:q] \;:=\; \sum_{i=1}^{k} D_{\mathrm{KL}}[(1-\alpha_i)p_{+}\alpha_i q : (1-\bar{\alpha})p + \bar{\alpha}q], \tag{6}$$

$$= \; h[(1-\bar{\alpha})p + \bar{\alpha}q] - \sum_{i=1}^{k} h[(1-\alpha_i)p_{+}\alpha_i q], \tag{7}$$

where $\bar{\alpha} = \sum_{i=1}^k w_i \alpha_i$. The divergence $D_{\mathrm{JS}}^{\alpha,w}$ generalizes the (scalar) skew Jensen-Shannon divergence when $k = 1$, and is a Ali-Silvey-Csiszár f-divergence upper bounded by $\log \frac{1}{\bar{\alpha}(1-\bar{\alpha})}$ [31].

- **A priori mid-density**. The JSD can be interpreted as the total divergence of the densities to the *mid-density* $\bar{p} = \sum_{i=1}^n w_i p_i$, a statistical mixture:

$$D_{\mathrm{JS}}[\mathcal{P}] = \sum_{i=1}^{n} w_i D_{\mathrm{KL}}[p_i : \bar{p}] = h[\bar{p}] - \sum_{i=1}^{n} w_i h[p_i].$$

Unfortunately, the JSD between two Gaussian densities is not known in closed form because of the definite integral of a log-sum term (i.e., K-divergence between a density and a mixture density $\bar{p}$). For the special case of the Cauchy family, a closed-form formula [41] for the JSD between two Cauchy densities was obtained. Thus we may choose a *geometric mixture distribution* [29] instead of the ordinary arithmetic mixture $\bar{p}$. More generally, we can choose any weighted mean M_α (say, the geometric mean, or the harmonic mean, or any other power mean) and define a generalization of the K-divergence of Equation 4:

$$D_K^{M_\alpha}[p:q] := D_K[p:(pq)_{M_\alpha}], \tag{8}$$

where

$$(pq)_{M_\alpha}(x) := \frac{M_\alpha(p(x), q(x))}{Z_{M_\alpha}(p:q)}$$

is a statistical M-mixture with $Z_{M_\alpha}(p,q)$ denoting the normalizing coefficient:

$$Z_{M_\alpha}(p:q) = \int M_\alpha(p(x), q(x))\mathrm{d}\mu(x)$$

so that $\int (pq)_{M_\alpha}(x)\mathrm{d}\mu(x) = 1$. These M-mixtures are well-defined provided the convergence of the definite integrals.

Then we define a generalization of the JSD [29] termed (M_α, N_β)*-Jensen-Shannon divergence* as follows:

$$D_{\mathrm{JS}}^{M_\alpha,N_\beta}[p:q] := N_\beta\left(D_K[p:(pq)_{M_\alpha}], D_K[q:(pq)_{M_\alpha}]\right), \tag{9}$$

where N_β is yet another weighted mean to average the two M_α-K-divergences. We have $D_{\mathrm{JS}} = D_{\mathrm{JS}}^{A,A}$ where $A(a,b) = \frac{a+b}{2}$ is the arithmetic mean. The geometric JSD yields a closed-form formula between two multivariate Gaussians, and has been used in deep learning [10]. More generally, we may consider the Jensen-Shannon symmetrization of an arbitrary distance D as

$$D^{\mathrm{JS}}_{M_\alpha,N_\beta}[p:q] := N_\beta\left(D[p:(pq)_{M_\alpha}], D[q:(pq)_{M_\alpha}]\right). \tag{10}$$

- **A posteriori mid-density**. We consider a generalization of Sibson's information radius [52]. Let $S_w(a_1, \ldots, a_n)$ denote a generic weighted mean of n positive scalars $a_1, \ldots, a_n$, with weight vector $w \in \Delta_n$. Then we define the *S-variational Jensen-Shannon diversity index* [34] as

$$D_{\mathrm{vJS}}^{S_w}(\mathcal{P}) := \min_c S_w \left(D_{\mathrm{KL}}[p_1 : c], D_{\mathrm{KL}}[p_n : c] \right). \tag{11}$$

When $S_w = A_w$ (with $A_w(a_1, \ldots, a_n) = \sum_{i=1}^n w_i a_i$ the arithmetic weighted mean), we recover the ordinary Jensen-Shannon diversity index. More generally, we define the *S-Jensen-Shannon index of an arbitrary distance D* as

$$D_{S_w}^{\mathrm{vJS}}(\mathcal{P}) := \min_c S_w \left(D[p_1 : c], \ldots, D[p_n : c] \right). \tag{12}$$

When $n = 2$, this yields a Jensen-Shannon-symmetrization of distance D.

The variational optimization defining the JSD can also be constrained to a (parametric) family of densities $\mathcal{D}$, thus defining the *$(S, \mathcal{D})$-relative Jensen-Shannon diversity index*:

$$D_{\mathrm{vJS}}^{S_w, \mathcal{D}}(\mathcal{P}) := \min_{c \in \mathcal{D}} S_w \left(D_{\mathrm{KL}}[p_1 : c], \ldots, D_{\mathrm{KL}}[p_n : c] \right). \tag{13}$$

The relative Jensen-Shannon divergences are useful for clustering applications: Let p_{θ_1} and p_{θ_2} be two densities of an exponential family $\mathcal{E}$ with cumulant function $F(\theta)$. Then the $\mathcal{E}$-relative Jensen-Shannon divergence is the Bregman information of $\mathcal{P}_2(p, q)$ for the conjugate function $F^*(\eta) = -h[p_\theta]$ (with $\eta = \nabla F(\theta)$). The $\mathcal{E}$-relative JSD amounts to a *Jensen divergence* for F^*:

$$\begin{aligned}
D_{\mathrm{vJS}}[p_{\theta_1}, p_{\theta_2}] &= \min_\theta \frac{1}{2} \left\{ D_{\mathrm{KL}}[p_{\theta_1} : p_\theta] + D_{\mathrm{KL}}[p_{\theta_2} : p_\theta] \right\}, && (14)\\
&= \min_\theta \frac{1}{2} \left\{ B_F[\theta : \theta_1] + B_F[\theta : \theta_2] \right\}, && (15)\\
&= \min_\eta \frac{1}{2} \left\{ B_{F^*}[\eta_1 : \eta] + B_{F^*}[\eta_2 : \eta] \right\}, && (16)\\
&= \frac{F^*(\eta_1) + F^*(\eta_2)}{2} - F^*(\eta^*), && (17)\\
&=: J_{F^*}(\eta_1, \eta_2), && (18)
\end{aligned}$$

since $\eta^* := \frac{\eta_1 + \eta_2}{2}$ (a right-sided *Bregman centroid* [36]).

7 Statistical distances between mixtures

Pearson [48] first considered a unimodal Gaussian mixture of two components for modeling distributions crabs in 1894. Statistical mixtures [25] like the Gaussian mixture models (GMMs) are often met in information sciences, and therefore it is important to assess their dissimilarities. Let $m(x) = \sum_{i=1}^k w_i p_i(x)$ and $m'(x) = \sum_{i=1}^{k'} w_i' p_i'(x)$ be two finite statistical mixtures. The KLD between two GMMs m and m' is not analytic [55] because of the log-sum terms:

$$D_{\mathrm{KL}}[m : m'] = \int m(x) \log \frac{m(x)}{m'(x)} \mathrm{d}x.$$

However, the KLD between two GMMs with the same prescribed components $p_i(x) = p_i'(x) = p_{\mu_i,\Sigma_i}(x)$ (i.e., $k = k'$, and only the normalized positive weights may differ) is provably a Bregman divergence [39] for the differential negentropy $F(\theta)$:

$$D_{\mathrm{KL}}[m(\theta) : m(\theta')] = B_F(\theta, \theta'),$$

where $m(\theta) = \sum_{i=1}^{k-1} w_i p_i(x) + (1 - \sum_{i=1}^{k-1} w_i) p_k(x)$ and $F(\theta) = \int m(\theta) \log m(\theta) \mathrm{d}x$. The family $\{m_\theta \ \ \theta \in \Delta_{k-1}^\circ\}$ is called a mixture family in information geometry, where Δ_{k-1}° denotes the $(k-1)$-dimensional open standard simplex. However, $F(\theta)$ is usually not available in closed-form because of the log-sum integral. In some special cases like the mixture of two prescribed Cauchy distributions, we get a closed-form formula for the KLD, JSD, etc. [41, 35]. Thus when dealing with mixtures (like GMMs), we either need efficient approximating (§7.1), bounding (§7.2) KLD techniques, or new distances (§7.3) that yields closed-form formula between mixture densities.

7.1 Approximating and/or fast statistical distances between mixtures

- The Jeffreys divergence (JD) $D_J[m, m'] = D_{\mathrm{KL}}[m : m'] + D_{\mathrm{KL}}[m' : m]$ between two (Gaussian) MMs is not available in closed-form, and can be estimated using Monte Carlo integration as

  $$\hat{D}_J^{\mathcal{S}_s}[m, m'] := \frac{1}{s} \sum_{i=1}^{s} 2 \frac{(m(x_i) - m'(x_i))}{m(x_i) + m'(x_i)} \log \left(\frac{m(x_i)}{m'(x_i)} \right),$$

  where $\mathcal{S}_s = \{x_1, \ldots, x_s\}$ are s IID samples from the mid mixture $m_{12}(x) := \frac{1}{2}(m(x) + m'(x))$ (with $\lim_{s\to\infty} \hat{D}_J^{\mathcal{S}_s}[m, m'] = D_J[m, m']$). In [33], the mixtures m and m' are converted into densities of an exponential-polynomial family. The JD between densities p_θ and $p_{\theta'}$ of an exponential family with cumulant function F is available in closed-form:

  $$D_J[p_\theta, p_{\theta'}] = (\theta' - \theta) \cdot (\eta' - \eta),$$

  with $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$, where F^* denotes the convex conjugate. Any smooth density r (includes a mixture $r = m$) is converted into close densities $p_{\theta_r^{\mathrm{MLE}}}$ and $p^{\eta_r^{\mathrm{SME}}}$ of a exponential-polynomial family using extensions of the Maximum Likelihood Estimator (MLE) and Score Matching Estimator (SME). Then JD between mixtures is approximated as follows

  $$D_J[m, m'] \simeq ({\theta'}^{\mathrm{SME}} - \theta^{\mathrm{SME}}) \cdot ({\eta'}^{\mathrm{MLE}} - \eta^{\mathrm{MLE}}).$$

- Given a finite set of mixtures $\{m_i(x)\}$ sharing the same components (e.g., points on a mixture family manifold), we precompute the KLD between pairwise components to obtain fast approximation of the KLD $D_{\mathrm{KL}}[m_i : m_j]$ between any two mixtures m_i and m_j, see [51].

7.2 Bounding statistical distances between mixtures

- **Log-Sum-Exp bounds**: In [42, 43], we lower and upper bound the cross-entropy between mixtures using the fact that the log-sum term $\log m(x)$ and be interpreted as a LSE function. We then compute lower envelopes and upper envelopes of density functions using technique of computational geometry to report deterministic lower and upper bounds on the KLD and α-divergences. These bounds are said combinatorial because we decompose the support into elementary intervals. Bounds between the Total Variation Distance (TVD) between univariate mixtures are reported in [44].

7.3 Newly designed statistical distances yielding closed-form formula for mixtures

- **Statistical Minkowski distances** [30]: Consider the Lebesgue space

$$L_\alpha(\mu) := \left\{ f \in \mathbb{F} \ : \ \int_{\mathcal{X}} |f(x)|^\alpha \mathrm{d}\mu(x) < \infty \right\}$$

  for $\alpha \geq 1$, where $\mathbb{F}$ denotes the set of all real-valued measurable functions defined on the support $\mathcal{X}$. Minkowski's inequality writes as $\|p+q\|_\alpha \leq \|p\|_\alpha + \|q\|_\alpha$ for $\alpha \in [1,\infty)$. The statistical Minkowski difference distance between $p, q \in L_\alpha(\mu)$ is defined as

$$D_\alpha^{\mathrm{Minkowski}}[p,q] := \|p\|_\alpha + \|q\|_\alpha - \|p+q\|_\alpha \geq 0. \quad (19)$$

  The statistical Minkowski log-ratio distance is defined by:

$$L_\alpha^{\mathrm{Minkowski}}[p,q] := -\log \frac{\|p+q\|_\alpha}{\|p\|_\alpha + \|q\|_\alpha} \geq 0. \quad (20)$$

  These statistical Minkowski distances are symmetric, and $L_\alpha[p,q]$ is scale-invariant. For even integers $\alpha \geq 2$, $D_\alpha^{\mathrm{Minkowski}}[m : m']$ is available in closed-form.

- We show that the Cauchy-Schwarz divergence (CSD), the quadratic Jensen-Rényi divergence [54] (JRD), and the total square Distance (TSD) between two GMMs, and more generally two mixtures of exponential families, can be obtained in closed-form in [28].

Initially created 13th August 2021 (last updated November 12, 2021).

References

[1] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[2] Shun-Ichi Amari. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

[3] Shun-ichi Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Japan, 2016.

[4] Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[5] Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical inference: the minimum distance approach*. Chapman and Hall/CRC, 2019.

[6] JM Corcuera and Federica Giummolè. A characterization of monotone and regular divergences. *Annals of the Institute of Statistical Mathematics*, 50(3):433–450, 1998.

(M,D)-information	Definition $I_{M,D}[p,q] := \min_c\{M(D[p:c],D[q:c])\}$
Information radius of order 1 aka Jensen-Shannon divergence	$I_{A,\mathrm{KL}}[p,q] = \frac{1}{2}D_{\mathrm{KL}}[p:\frac{p+q}{2}] + \frac{1}{2}D_{\mathrm{KL}}[q:\frac{p+q}{2}]$
Information radius of order α aka Sibson's information radius	$I_{M_\alpha^R,D_\alpha^R}[p,q] = \frac{\alpha}{\alpha-1}\log_2\int_{\mathcal{X}}\left(\frac{p(x)^\alpha+q(x)^\alpha}{2}\right)^{\frac{1}{\alpha}}\mathrm{d}\mu(x)$
Bregman information	$I_{A,B_F}(\theta_1,\theta_2) = \frac{1}{2}B_F(\theta_1:\frac{\theta_1+\theta_2}{2}) + \frac{1}{2}B_F(\theta_2:\frac{\theta_1+\theta_2}{2})$
skewed Jensen–Bregman divergence	$I_{A_\beta,B_F}(\theta_1:\theta_2) = (1-\beta)B_F(\theta_1:(1-\beta)\theta_1+\beta\theta_2) + \beta B_F(\theta_2:(1-\beta)\theta_1 +$
Chernoff minimax discrimination	$I_{\max,D_{\mathrm{KL}}^*}[p,q] = \min_c\max\{D_{\mathrm{KL}}[c:p], D_{\mathrm{KL}}[c:q]\}$
Amari's α-risk	$I_{A,D_\alpha}[p,q] = \frac{1}{2}D_\alpha[p:(pq)_\alpha^A] + \frac{1}{2}D_\alpha[q:(pq)_\alpha^A]$
Annealing geometric paths $(pq)_{G_\beta}$	minimizer of $I_{G_\beta,D_{\mathrm{KL}}^*}[p:q]$

arithmetic mean	$A(a,b) = \frac{a+b}{2}$
weighted arithmetic mean	$A_\beta(a,b) = (1-\beta)a + \beta b$
geometric mean	$G_\beta(a,b) = a^{1-\beta}b^\beta$
maximum (mean)	$\mathrm{MAX}(a,b) = \max\{a,b\}$
Rényi's α-mean	$M_\alpha^R(a,b) = \frac{1}{\alpha-1}\log_2\left(\frac{2^{(\alpha-1)a}+2^{(\alpha-1)b}}{2}\right)$
Amari's α-integration	$(pq)_\alpha^A(x) \propto f_\alpha^{-1}(\frac{1}{2}f_\alpha(p(x)) + \frac{1}{2}f_\alpha(q(x)))$, α-representation $f_\alpha(u) = \frac{2}{1-\alpha}u^{\frac{1-\alpha}{2}}$
M-mixture	$(pq)_{M_\beta}(x) \propto M_\beta(p(x),q(x))$

Kullback-Leibler divergence	$D_{\mathrm{KL}}[p:q] = \int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x)$
reverse Kullback-Leibler divergence	$D_{\mathrm{KL}}^*[p:q] = \int_{\mathcal{X}} q(x)\log\frac{q(x)}{p(x)}\mathrm{d}\mu(x) = D_{\mathrm{KL}}[q:p]$
Rényi's α-divergence	$D_\alpha^R[p:q] = \frac{1}{\alpha-1}\log_2\left(\int_{\mathcal{X}} p(x)^\alpha q(x)^{1-\alpha}\mathrm{d}\mu(x)\right)$
Amari's α-divergence	$D_\alpha^A[p:q] = \frac{4}{1-\alpha^2}(1-\int_{\mathcal{X}} p(x)^{\frac{1-\alpha}{2}}q(x)^{\frac{1+\alpha}{2}}\mathrm{d}\mu(x))$

Table 1: Examples of information radius measures as variational abstract mean divergences.

[7] Imre Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

[8] Imre Csiszár. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4):191–213, 1972.

[9] Imre Csiszár and Paul C Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.

[10] Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Constraining Variational Inference with Geometric Jensen-Shannon Divergence. In *Advances in Neural Information Processing Systems*, 2020.

[11] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[12] Bent Fuglede and Flemming Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE, 2004.

[13] Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

[14] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

[15] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[16] Robert Jenssen, Jose C Principe, Deniz Erdogmus, and Torbjørn Eltoft. The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *Journal of the Franklin Institute*, 343(6):614–629, 2006.

[17] Jiantao Jiao, Thomas A Courtade, Albert No, Kartik Venkat, and Tsachy Weissman. Information measures: the curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60(12):7616–7626, 2014.

[18] MC Jones, Nils Lid Hjort, Ian R Harris, and Ayanendranath Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.

[19] Kittipat Kampa, Erion Hasanbelliu, and Jose C Principe. Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians. In *The 2011 International Joint Conference on Neural Networks*, pages 2578–2585. IEEE, 2011.

[20] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[21] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics (AISTATS)*, page 65?72, 2001.

[22] Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[23] Jianhua Lin and SKM Wong. Approximation of discrete probability distributions based on a new divergence measure. *Congressus Numerantium (Winnipeg)*, 61:75–80, 1988.

[24] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[25] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

[26] Tetsuzo Morimoto. Markov processes and the H-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.

[27] Frank Nielsen. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv preprint arXiv:1009.4004*, 2010.

[28] Frank Nielsen. Closed-form information-theoretic divergences for statistical mixtures. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR)*, pages 1723–1726. IEEE, 2012.

[29] Frank Nielsen. On the Jensen?Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy*, 21(5), 2019.

[30] Frank Nielsen. The statistical Minkowski distances: Closed-form formula for Gaussian mixture models. In *International Conference on Geometric Science of Information*, pages 359–367. Springer, 2019.

[31] Frank Nielsen. On a Generalization of the Jensen?Shannon Divergence and the Jensen?Shannon Centroid. *Entropy*, 22(2), 2020.

[32] Frank Nielsen. Fast approximations of the Jeffreys divergence between univariate Gaussian mixture models via exponential polynomial densities. *arXiv preprint arXiv:2107.05901*, 2021.

[33] Frank Nielsen. Fast approximations of the jeffreys divergence between univariate gaussian mixture models via exponential polynomial densities. *arXiv preprint arXiv:2107.05901*, 2021.

[34] Frank Nielsen. On a Variational Definition for the Jensen-Shannon Symmetrization of Distances Based on the Information Radius. *Entropy*, 23(4), 2021.

[35] Frank Nielsen. The dually flat information geometry of the mixture family of two prescribed Cauchy components. *arXiv preprint arXiv:2104.13801*, 2021.

[36] Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

[37] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2013.

[38] Frank Nielsen and Richard Nock. Patch matching with polynomial exponential families and projective divergences. In *International Conference on Similarity Search and Applications*, pages 109–116. Springer, 2016.

[39] Frank Nielsen and Richard Nock. On the geometry of mixtures of prescribed distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2861–2865. IEEE, 2018.

[40] Frank Nielsen and Kazuki Okamura. On f-divergences between Cauchy distributions. *arXiv preprint arXiv:2101.12459*, 2021.

[41] Frank Nielsen and Kazuki Okamura. On f-divergences between cauchy distributions. *arXiv:2101.12459*, 2021.

[42] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016.

[43] Frank Nielsen and Ke Sun. Combinatorial bounds on the α-divergence of univariate mixture models. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 4476–4480. IEEE, 2017.

[44] Frank Nielsen and Ke Sun. Guaranteed Deterministic Bounds on the total variation distance between univariate mixtures. In *28th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2018, Aalborg, Denmark, September 17-20, 2018*, pages 1–6. IEEE, 2018.

[45] Frank Nielsen and Ke Sun. Clustering in Hilbert's projective geometry: The case studies of the probability simplex and the elliptope of correlation matrices. In *Geometric Structures of Information*, pages 297–331. Springer, 2019.

[46] Frank Nielsen, Ke Sun, and Stéphane Marchand-Maillet. On Hölder projective divergences. *Entropy*, 19(3):122, 2017.

[47] Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[48] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[49] Souvik Ray, Subrata Pal, Sumit Kumar Kar, and Ayanendranath Basu. Characterizing the functional density power divergence class. *arXiv preprint arXiv:2105.06094*, 2021.

[50] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[51] Olivier Schwander, Stéphane Marchand-Maillet, and Frank Nielsen. Comix: Joint estimation and lightspeed comparison of mixture models. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, pages 2449–2453. IEEE, 2016.

[52] Robin Sibson. Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(2):149–160, 1969.

[53] Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18, 2017.

[54] Fei Wang, Tanveer Syeda-Mahmood, Baba C Vemuri, David Beymer, and Anand Rangarajan. Closed-form Jensen-Rényi divergence for mixture of Gaussians and applications to group-wise shape registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–655. Springer, 2009.

[55] Sumio Watanabe, Keisuke Yamazaki, and Miki Aoyagi. Kullback information of normal mixture is not an analytic function. *IEICE technical report. Neurocomputing*, 104(225):41–46, 2004.

[56] Andrew KC Wong and Manlai You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):599–609, 1985.