entropy

MDPI

Article

On Hölder Projective Divergences

Frank Nielsen [1,2,*], Ke Sun [3] and Stéphane Marchand-Maillet [4]

1 Computer Science Department LIX, École Polytechnique, 91128 Palaiseau Cedex, France
2 Sony Computer Science Laboratories Inc., Tokyo 141-0022, Japan
3 Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia; sunk@ieee.org
4 Computer Vision and Multimedia Laboratory (Viper), University of Geneva, CH-1211 Geneva, Switzerland; Stephane.Marchand-Maillet@unige.ch
* Correspondence: Frank.Nielsen@acm.org or nielsen@lix.polytechnique.fr; Tel.: +33-1-7757-8070

Academic Editor: Geert Verdoolaege
Received: 20 January 2017; Accepted: 10 March 2017; Published: 16 March 2017

Abstract: We describe a framework to build distances by measuring the tightness of inequalities and introduce the notion of proper statistical divergences and improper pseudo-divergences. We then consider the Hölder ordinary and reverse inequalities and present two novel classes of Hölder divergences and pseudo-divergences that both encapsulate the special case of the Cauchy–Schwarz divergence. We report closed-form formulas for those statistical dissimilarities when considering distributions belonging to the same exponential family provided that the natural parameter space is a cone (e.g., multivariate Gaussians) or affine (e.g., categorical distributions). Those new classes of Hölder distances are invariant to rescaling and thus do not require distributions to be normalized. Finally, we show how to compute statistical Hölder centroids with respect to those divergences and carry out center-based clustering toy experiments on a set of Gaussian distributions which demonstrate empirically that symmetrized Hölder divergences outperform the symmetric Cauchy–Schwarz divergence.

Keywords: Hölder inequalities; Hölder divergences; projective divergences; Cauchy–Schwarz divergence; Hölder escort divergences; skew Bhattacharyya divergences; exponential families; conic exponential families; escort distribution; clustering

1. Introduction: Inequality, Proper Divergence and Improper Pseudo-Divergence

1.1. Statistical Divergences from Inequality Gaps

An inequality [1] is denoted mathematically by $\text{lhs} \leq \text{rhs}$, where lhs and rhs denote respectively the left-hand-side and right-hand-side of the inequality. One can build dissimilarity measures from inequalities $\text{lhs} \leq \text{rhs}$ by measuring the inequality tightness: For example, we may quantify the tightness of an inequality by its difference gap:

$$\Delta = \text{rhs} - \text{lhs} \geq 0. \tag{1}$$

When $\text{lhs} > 0$, the inequality tightness can also be gauged by the log-ratio gap:

$$D = \log\left(\frac{\text{rhs}}{\text{lhs}}\right) = -\log\left(\frac{\text{lhs}}{\text{rhs}}\right) \geq 0. \tag{2}$$

We may further compose this inequality tightness value measuring non-negative gaps with a strictly monotonically increasing function f (with $f(0) = 0$).

A bi-parametric inequality $\mathrm{lhs}(p,q) \leq \mathrm{rhs}(p,q)$ is called proper if it is strict for $p \neq q$ (i.e., $\mathrm{lhs}(p,q) < \mathrm{rhs}(p,q), \forall p \neq q$) and tight if and only if (iff) $p = q$ (i.e., $\mathrm{lhs}(p,q) = \mathrm{rhs}(p,q), \forall p = q$). Thus a proper bi-parametric inequality allows one to define dissimilarities such that $D(p,q) = 0$ iff $p = q$. Such a dissimilarity is called proper. Otherwise, an inequality or dissimilarity is said to be improper. Note that there are many equivalent words used in the literature instead of (dis-)similarity: distance (although often assumed to have metric properties; here, we used the notion of distance as a dissimilarity that may be asymmetric), pseudo-distance, discrimination, proximity, information deviation, etc.

A statistical dissimilarity between two discrete or continuous distributions $p(x)$ and $q(x)$ on a support $\mathcal{X}$ can thus be defined from inequalities by summing up or taking the integral for the inequalities instantiated on the observation space $\mathcal{X}$:

$$\begin{aligned} &\forall x \in \mathcal{X}, \quad D_x(p,q) = \mathrm{rhs}(p(x),q(x)) - \mathrm{lhs}(p(x),q(x)) \Rightarrow \\ &\qquad D(p,q) = \begin{cases} \sum_{x \in \mathcal{X}} \left[\mathrm{rhs}(p(x),q(x)) - \mathrm{lhs}(p(x),q(x))\right] & \text{discrete case,} \\ \int_{\mathcal{X}} \left[\mathrm{rhs}(p(x),q(x)) - \mathrm{lhs}(p(x),q(x))\right] \mathrm{d}x & \text{continuous case.} \end{cases} \end{aligned} \tag{3}$$

In such a case, we get a separable divergence by construction. Some non-separable inequalities induce a non-separable divergence. For example, the renowned Cauchy–Schwarz divergence [2] is not separable because in the inequality:

$$\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x \leq \sqrt{\left(\int_{\mathcal{X}} p(x)^2 \mathrm{d}x\right)\left(\int_{\mathcal{X}} q(x)^2 \mathrm{d}x\right)}, \tag{4}$$

the rhs is not separable.

Furthermore, a proper dissimilarity is called a divergence in information geometry [3] when it is C^3 (i.e., three times differentiable, thus allowing one to define a metric tensor [4] and a cubic tensor [3]).

Many familiar distances can be reinterpreted as inequality gaps in disguise. For example, Bregman divergences [5] and Jensen divergences [6] (also called Burbea–Rao divergences [7,8]) can be reinterpreted as inequality difference gaps and the Cauchy–Schwarz distance [2] as an inequality log-ratio gap:

Example 1 (Bregman divergence as a Bregman score-induced gap divergence)**.** *A proper score function [9] $S(p:q)$ induces a gap divergence $D(p:q) = S(p:q) - S(p:p) \geq 0$. A Bregman divergence [5] $B_F(p:q)$ for a strictly convex and differentiable real-valued generator $F(x)$ is induced by the Bregman score $S_F(p:q)$. Let $S_F(p:q) = -F(q) - \langle p - q, \nabla F(q)\rangle$ denote the Bregman proper score minimized for $p = q$. Then, the Bregman divergence is a gap divergence: $B_F(p:q) = S_F(p:q) - S_F(p:p) \geq 0$. When F is strictly convex, the Bregman score is proper, and the Bregman divergence is proper.*

Example 2 (Cauchy–Schwarz distance as a log-ratio gap divergence)**.** *Consider the Cauchy–Schwarz inequality $\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x \leq \sqrt{\left(\int_{\mathcal{X}} p(x)^2\mathrm{d}x\right)\left(\int_{\mathcal{X}} q(x)^2\mathrm{d}x\right)}$. Then, the Cauchy–Schwarz distance [2] between two continuous distributions is defined by* $\mathrm{CS}(p:q) = -\log \frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x}{\sqrt{\left(\int_{\mathcal{X}} p(x)^2\mathrm{d}x\right)\left(\int_{\mathcal{X}} q(x)^2\mathrm{d}x\right)}} \geq 0$.

Note that we use the modern notation $D(p:q)$ to emphasize that the divergence is potentially asymmetric: $D(p:q) \neq D(q:p)$; see [3]. In information theory [10], the older notation "$\|$" is often used instead of the ":" that is used in information geometry [3].

To conclude this introduction, let us finally introduce the notion of projective statistical distances. A statistical distance $D(p:q)$ is said to be projective when it is invariant to scaling of $p(x)$ and $q(x)$, that is,

$$D(\lambda p : \lambda' q) = D(p:q), \quad \forall \lambda, \lambda' > 0. \tag{5}$$

The Cauchy–Schwarz distance is a projective divergence. Another example of such a projective divergence is the parametric γ-divergence [11].

Example 3 (γ-divergence as a projective score-induced gap divergence)**.** *The γ-divergence [11,12] $D_\gamma(p:\ q)$ for $\gamma > 0$ is projective:*

$$D_\gamma(p:q) = S_\gamma(p:q) - S_\gamma(p:p), \text{ with}$$
$$S_\gamma(p:q) = -\frac{1}{\gamma(1+\gamma)} \frac{\int p(x)q(x)^\gamma \mathrm{d}x}{(\int q(x)^{1+\gamma}\mathrm{d}x)^{\frac{\gamma}{1+\gamma}}}.$$

The γ-divergence is related to the proper pseudo-spherical score [11].

The γ-divergences have been proven useful for robust statistical inference [11] in the presence of heavy outlier contamination. In general, bi-parametric homogeneous inequalities yield corresponding log-ratio projective divergences: Let $\mathrm{lhs}(p:q)$ and $\mathrm{rhs}(p:q)$ be homogeneous functions of degree $k \in \mathbb{N}$ (i.e., $\mathrm{lhs}(\lambda p:\lambda' q) = (\lambda\lambda')^k \mathrm{lhs}(p:q)$ and $\mathrm{rhs}(\lambda p:\lambda' q) = (\lambda\lambda')^k \mathrm{rhs}(p:q)$); then, it comes that:

$$D(\lambda p:\lambda' q) = -\log\left(\frac{\mathrm{lhs}(\lambda p:\lambda' q)}{\mathrm{rhs}(\lambda p:\lambda' q)}\right) = -\log\left(\frac{(\lambda\lambda')^k \mathrm{lhs}(p:q)}{(\lambda\lambda')^k \mathrm{rhs}(p:q)}\right) = -\log\left(\frac{\mathrm{lhs}(p:q)}{\mathrm{rhs}(p:q)}\right) = D(p:q). \quad (6)$$

For example, Hölder and Cauchy–Schwarz inequalities are homogeneous inequalities of degree one that yield projective log-ratio divergences.

There are many works studying classes of (statistical) divergences and their properties. For example, Zhang [13] studied the relationships between divergences, duality and convex analysis by defining the class of divergences:

$$D_F^{(\alpha)}(p:q) = \frac{4}{1-\alpha^2}\left(\frac{1-\alpha}{2}F(p) + \frac{1+\alpha}{2}F(q) - F\left(\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q\right)\right), \alpha \neq 1, \quad (7)$$

for a real-valued convex generator function F. Interestingly, this divergence can be interpreted as a gap divergence derived from the Jensen convex inequality:

$$\frac{1-\alpha}{2}F(p) + \frac{1+\alpha}{2}F(q) \geq F\left(\frac{1-\alpha}{2}p + \frac{1+\alpha}{2}q\right). \quad (8)$$

This work is further extended in [14] where Zhang stresses the two different types of duality in information geometry: the referential duality and the representational duality (with the study of the (ρ,τ)-geometry for monotone embeddings).

It is well-known that Rényi divergence generalizes the Kullback–Leibler divergence: Rényi divergence is induced by Rényi entropy, which generalizes Shannon entropy, while keeping the important feature of being additive. Another generalization of Shannon entropy is Tsallis entropy, which is non-additive in general and allows one to define the Tsallis divergence. Both the Rényi and Tsallis entropies can be unified by the biparametric family of Sharma–Mittal entropies [15], and the corresponding Sharma–Mittal divergences can be defined. There are many ways to extend the definitions of Sharma–Mittal divergences. For example, in [16], a generalization of Rényi divergences is proposed, and its induced geometry is investigated.

1.2. Pseudo-Divergences and the Axiom of Indiscernibility

Consider a broader class of statistical pseudo-divergences based on improper inequalities, where the tightness of an inequality $\mathrm{lhs}(p,q) \leq \mathrm{rhs}(p,q)$ does not imply that $p = q$. This family of dissimilarity measures has interesting properties that have not been studied before.

Formally, statistical pseudo-divergences are defined with respect to density measures $p(x)$ and $q(x)$ with $x \in \mathcal{X}$, where $\mathcal{X}$ denotes the support. By definition, pseudo-divergences satisfy the following three fundamental properties:

1. Non-negativeness: $D(p:q) \geq 0$ for any $p(x), q(x)$;
2. Reachable indiscernibility:
   - $\forall p(x)$, there exists $q(x)$, such that $D(p:q) = 0$,
   - $\forall q(x)$, there exists $p(x)$, such that $D(p:q) = 0$.
3. Positive correlation: if $D(p:q) = 0$, then $(p(x_1) - p(x_2))\,(q(x_1) - q(x_2)) \geq 0$ for any $x_1, x_2 \in \mathcal{X}$.

As compared to statistical divergence measures, such as the Kullback–Leibler (KL) divergence:

$$\mathrm{KL}(p:q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x, \tag{9}$$

pseudo-divergences do not require $D(p:p) = 0$. Instead, any pair of distributions $p(x)$ and $q(x)$ with $D(p:q) = 0$ only have to be "positively correlated" such that $p(x_1) \leq p(x_2)$ implies $q(x_1) \leq q(x_2)$, and vice versa. Any divergence with $D(p:q) = 0 \Rightarrow p(x) = q(x)$ (law of indiscernibles) automatically satisfies this weaker condition, and therefore, any divergence belongs to the broader class of pseudo-divergences. Indeed, if $p(x) = q(x)$, then $(p(x_1) - p(x_2))(q(x_1) - q(x_2)) = (p(x_1) - p(x_2))^2 \geq 0$. However, the converse is not true. As we shall describe in the remainder, the family of pseudo-divergences is not limited to proper divergence measures. In the remainder, the term "pseudo-divergence" refers to such divergences that are not proper divergence measures.

We study two novel statistical dissimilarity families: one family of statistical improper pseudo-divergences and one family of proper statistical divergences. Within the class of pseudo-divergences, this work concentrates on defining a tri-parametric family of dissimilarities called Hölder log-ratio gap divergence that we concisely abbreviate as HPD for "Hölder pseudo divergence" in the remainder. We also study its proper divergence counterpart termed HD for "Hölder divergence".

1.3. Prior Work and Contributions

The term "Hölder divergence" was first coined in 2014 based on the definition of the Hölder score [17,18]: The score-induced Hölder divergence $D(p:q)$ is a proper gap divergence that yields a scale-invariant divergence. Let $p_{a,\sigma}(x) = a\sigma p(\sigma x)$ for $a, \sigma > 0$ be a transformation. Then, a scale-invariant divergence satisfies $D(p_{a,\sigma} : q_{a,\sigma}) = \kappa(a,\sigma) D(p:q)$ for a function $\kappa(a,\sigma) > 0$. This gap divergence is proper since it is based on the so-called Hölder score, but is not projective and does not include the Cauchy–Schwarz divergence. Due to these differences, the Hölder log-ratio gap divergence introduced here shall not be confused with the Hölder gap divergence induced by the Hölder score that relies both on a scalar γ and a function $\phi(\cdot)$.

We shall introduce two novel families of log-ratio projective gap divergences based on Hölder ordinary (or forward) and reverse inequalities that extend the Cauchy–Schwarz divergence, study their properties and consider as an application clustering Gaussian distributions: We experimentally show better clustering results when using symmetrized Hölder divergences than using the Cauchy–Schwarz divergence. To contrast with the "Hölder composite score-induced divergences" of [18], our Hölder divergences admit closed-form expressions between distributions belonging to the same exponential families [19] provided that the natural parameter space is a cone or affine.

Our main contributions are summarized as follows:

- Define the tri-parametric family of Hölder improper pseudo-divergences (HPDs) in Section 2 and the bi-parametric family of Hölder proper divergences in Section 3 (HDs) for positive and probability measures, and study their properties (including their relationships with skewed Bhattacharyya distances [8] via escort distributions);

- Report closed-form expressions of those divergences for exponential families when the natural parameter space is a cone or affine (including, but not limited to the cases of categorical distributions and multivariate Gaussian distributions) in Section 4;
- Provide approximation techniques to compute those divergences between mixtures based on log-sum-exp inequalities in Section 4.6;
- Describe a variational center-based clustering technique based on the convex-concave procedure for computing Hölder centroids, and report our experimental results in Section 5.

1.4. Organization

This paper is organized as follows: Section 2 introduces the definition and properties of Hölder pseudo-divergences (HPDs). It is followed by Section 3 that describes Hölder proper divergences (HDs). In Section 4, closed-form expressions for those novel families of divergences are reported for the categorical, multivariate Gaussian, Bernoulli, Laplace and Wishart distributions. Section 5 defines Hölder statistical centroids and presents a variational k-means clustering technique: we show experimentally that using Hölder divergences improves clustering quality over the Cauchy–Schwarz divergence. Finally, Section 6 concludes this work and hints at further perspectives from the viewpoint of statistical estimation and manifold learning. In Appendix A, we recall the proof of the ordinary and reverse Hölder's inequalities.

2. Hölder Pseudo-Divergence: Definition and Properties

Hölder's inequality (see [20,21] and Appendix A for a proof) states for positive real-valued functions $p(x)$ and $q(x)$ defined on the support $\mathcal{X}$ that:

$$\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x \leq \left(\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x\right)^{\frac{1}{\alpha}} \left(\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x\right)^{\frac{1}{\beta}}, \tag{10}$$

where exponents α and β satisfy $\alpha\beta > 0$, as well as the exponent conjugacy condition: $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. In a more general form, Hölder's inequality holds for any real and complex valued functions. In this work, we only focus on real positive functions that are densities of positive measures. We also write $\beta = \bar{\alpha} = \frac{\alpha}{\alpha-1}$, meaning that α and β are conjugate Hölder exponents. We check that $\alpha > 1$ and $\beta > 1$. Hölder inequality holds even if the lhs is infinite (meaning that the integral diverges), since the rhs is also infinite in that case.

The reverse Hölder inequality holds for conjugate exponents $\frac{1}{\alpha} + \frac{1}{\beta} = 1$ with $\alpha\beta < 0$ (then $0 < \alpha < 1$ and $\beta < 0$, or $\alpha < 0$ and $0 < \beta < 1$):

$$\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x \geq \left(\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x\right)^{\frac{1}{\alpha}} \left(\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x\right)^{\frac{1}{\beta}}. \tag{11}$$

Both Hölder's inequality and the reverse Hölder inequality turn tight when $p(x)^{\alpha} \propto q(x)^{\beta}$ (see proof in Appendix A).

2.1. Definition

Let $(\mathcal{X}, \mathcal{F}, \mu)$ be a measurable space where μ is the Lebesgue measure, and let $L^{\gamma}(\mathcal{X}, \mu)$ denote the Lebesgue space of functions that have their γ-th power of absolute value Lebesgue integrable, for any $\gamma > 0$ (when $\gamma \geq 1$, $L^{\gamma}(\mathcal{X}, \mu)$ is a Banach space). We define the following pseudo-divergence:

Definition 1 (Hölder statistical pseudo-divergence)**.** *For conjugate exponents α and β with $\alpha\beta > 0$ and $\sigma, \tau > 0$, the Hölder pseudo-divergence (HPD) between two densities $p(x) \in L^{\alpha\sigma}(\mathcal{X}, \mu)$ and $q(x) \in L^{\beta\tau}(\mathcal{X}, \mu)$ of positive measures absolutely continuous with respect to (w.r.t.) μ is defined by the following log-ratio gap:*

$$D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) = -\log\left(\frac{\int_{\mathcal{X}} p(x)^{\sigma} q(x)^{\tau}\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int_{\mathcal{X}} q(x)^{\beta\tau}\mathrm{d}x\right)^{\frac{1}{\beta}}}\right). \tag{12}$$

When $0 < \alpha < 1$ and $\beta = \bar{\alpha} = \frac{\alpha}{\alpha-1} < 0$, or $\alpha < 0$ and $0 < \beta < 1$, and $\sigma, \tau > 0$, the reverse HPD is defined by:

$$D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) = \log\left(\frac{\int_{\mathcal{X}} p(x)^{\sigma} q(x)^{\tau}\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int_{\mathcal{X}} q(x)^{\beta\tau}\mathrm{d}x\right)^{\frac{1}{\beta}}}\right). \tag{13}$$

By Hölder's inequality and the reverse Hölder inequality, $D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) \geq 0$ with $D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) = 0$ iff $p(x)^{\alpha\sigma} \propto q(x)^{\beta\tau}$ or equivalently $q(x) \propto p(x)^{\frac{\alpha\sigma}{\beta\tau}} = p(x)^{\frac{\sigma}{\tau}(\alpha-1)}$. When $\alpha > 1$, $x^{\frac{\sigma}{\tau}(\alpha-1)}$ is monotonically increasing, and $D^{\mathtt{H}}_{\alpha,\sigma,\tau}$ is indeed a pseudo-divergence. However, the reverse HPD is not a pseudo-divergence because $x^{\frac{\sigma}{\tau}(\alpha-1)}$ will be monotonically decreasing if $\alpha < 0$ or $0 < \alpha < 1$. Therefore, we only consider HPD with $\alpha > 1$ in the remainder, and leave here the notion of reverse Hölder divergence for future studies.

When $\alpha = \beta = 2$, $\sigma = \tau = 1$, the HPD becomes the Cauchy–Schwarz divergence CS [22]:

$$D^{\mathtt{H}}_{2,1,1}(p:q) = \mathrm{CS}(p:q) = -\log\left(\frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^2\mathrm{d}x\right)^{\frac{1}{2}}\left(\int_{\mathcal{X}} q(x)^2\mathrm{d}x\right)^{\frac{1}{2}}}\right), \tag{14}$$

which has been proven useful to get closed-form divergence formulas between mixtures of exponential families with conic or affine natural parameter spaces [23].

The Cauchy–Schwarz divergence is proper for probability densities since the Cauchy–Schwarz inequality becomes an equality iff $q(x) = \lambda p(x)^{\frac{\sigma}{\tau}(\alpha-1)} = \lambda p(x)$ implying that $\lambda = \int_{\mathcal{X}} \lambda p(x)\mathrm{d}x = \int_{\mathcal{X}} q(x)\mathrm{d}x = 1$. It is however not proper for positive densities.

Fact 1 (CS is only proper for probability densities)**.** *The Cauchy–Schwarz divergence $\mathrm{CS}(p:q)$ is proper for square-integrable probability densities $p(x), q(x) \in L^2(\mathcal{X}, \mu)$, but not proper for positive square-integrable densities.*

2.2. Properness and Improperness

In the general case, the divergence $D^{\mathtt{H}}_{\alpha,\sigma,\tau}$ is not even proper for normalized (probability) densities, not to mention general unnormalized (positive) densities. Indeed, when $p(x) = q(x)$, we have :

$$D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:p) = -\log\left(\frac{\int p(x)^{\sigma+\tau}\mathrm{d}x}{\left(\int p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int p(x)^{\beta\tau}\mathrm{d}x\right)^{\frac{1}{\beta}}}\right) \neq 0 \text{ when } \alpha\sigma \neq \beta\tau. \tag{15}$$

Let us consider the general case. For unnormalized positive distributions $\tilde{p}(x)$ and $\tilde{q}(x)$ (the tilde notation stems from the notation of homogeneous coordinates in projective geometry), the inequality becomes an equality when: $\tilde{p}(x)^{\alpha\sigma} \propto \tilde{q}(x)^{\beta\tau}$, i.e., $p(x)^{\alpha\sigma} \propto q(x)^{\beta\tau}$, or $q(x) \propto p(x)^{\alpha\sigma/\bar{\alpha}\tau} = p(x)^{\frac{\sigma}{\tau}(\alpha-1)}$. We can check that $D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:\lambda p^{\frac{\sigma}{\tau}(\alpha-1)}) = 0$ for any $\lambda > 0$:

$$-\log\left(\frac{\int p(x)^{\sigma}\lambda^{\tau}p(x)^{\sigma(\alpha-1)}\mathrm{d}x}{\left(\int p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int \lambda^{\beta\tau}p(x)^{(\alpha-1)\beta\sigma}\mathrm{d}x\right)^{\frac{1}{\beta}}}\right) = -\log\left(\frac{\int \lambda^{\tau}p(x)^{\alpha\sigma}\mathrm{d}x}{\left(\int p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int \lambda^{\beta\tau}p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\beta}}}\right) = 0, \tag{16}$$

since $(\alpha-1)\beta = (\alpha-1)\bar{\alpha} = (\alpha-1)\frac{\alpha}{\alpha-1} = \alpha$.

Fact 2 (HPD is improper)**.** *The Hölder pseudo-divergences are improper statistical distances.*

2.3. Reference Duality

In general, Hölder divergences are asymmetric when $\alpha \neq \beta$ ($\neq 2$) or $\sigma \neq \tau$, but enjoy the following reference duality [24]:

$$D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) = D^{\mathrm{H}}_{\beta,\tau,\sigma}(q:p) = D^{\mathrm{H}}_{\frac{\alpha}{\alpha-1},\tau,\sigma}(q:p). \tag{17}$$

Fact 3 (Reference duality HPD)**.** *The Hölder pseudo-divergences satisfy the reference duality* $\beta = \bar{\alpha} = \frac{\alpha}{\alpha-1}$*:* $D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) = D^{\mathrm{H}}_{\beta,\tau,\sigma}(q:p) = D^{\mathrm{H}}_{\frac{\alpha}{\alpha-1},\tau,\sigma}(q:p)$.

An arithmetic symmetrization of the HPD yields a symmetric HPD $S^{\mathrm{H}}_{\alpha,\sigma,\tau}$, given by:

$$\begin{aligned} S^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) &= S^{\mathrm{H}}_{\alpha,\sigma,\tau}(q:p) = \frac{D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) + D^{\mathrm{H}}_{\alpha,\sigma,\tau}(q:p)}{2}, \\ &= -\frac{1}{2}\log\left(\frac{\int p(x)^{\sigma}q(x)^{\tau}\mathrm{d}x \int p(x)^{\tau}q(x)^{\sigma}\mathrm{d}x}{\left(\int p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int p(x)^{\beta\tau}\mathrm{d}x\right)^{\frac{1}{\beta}}\left(\int q(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int q(x)^{\beta\tau}\mathrm{d}x\right)^{\frac{1}{\beta}}}\right). \end{aligned} \tag{18}$$

2.4. HPD is a Projective Divergence

In the above definition, densities $p(x)$ and $q(x)$ can either be positive or normalized probability distributions. Let $\tilde{p}(x)$ and $\tilde{q}(x)$ denote positive (not necessarily normalized) measures, and $w(\tilde{p}) = \int_{\mathcal{X}} \tilde{p}(x)\mathrm{d}x$ the overall mass so that $p(x) = \frac{\tilde{p}(x)}{w(\tilde{p})}$ is the corresponding normalized probability measure. Then, we check that HPD is a projective divergence [11] since:

$$D^{\mathrm{H}}_{\alpha,\sigma,\tau}(\tilde{p}:\tilde{q}) = D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q), \tag{19}$$

or in general:

$$D^{\mathrm{H}}_{\alpha,\sigma,\tau}(\lambda p:\lambda' q) = D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) \tag{20}$$

for all prescribed constants $\lambda, \lambda' > 0$. Projective divergences may also be called "angular divergences" or "cosine divergences", since they do not depend on the total mass of the density measures.

Fact 4 (HPD is projective)**.** *The Hölder pseudo-divergences are projective distances.*

2.5. Escort Distributions and Skew Bhattacharyya Divergences

Let us define with respect to the probability measures $p(x) \in L^{\frac{1}{\alpha}}(\mathcal{X},\mu)$ and $q(x) \in L^{\frac{1}{\beta}}(\mathcal{X},\mu)$ the following escort probability distributions [3]:

$$p^E_{\alpha}(x) = \frac{p(x)^{\frac{1}{\alpha}}}{\int p(x)^{\frac{1}{\alpha}}\mathrm{d}x}, \tag{21}$$

and

$$q^E_{\beta}(x) = \frac{q(x)^{\frac{1}{\beta}}}{\int q(x)^{\frac{1}{\beta}}\mathrm{d}x}. \tag{22}$$

Since HPD is a projective divergence, we compute with respect to the conjugate exponents α and β the Hölder escort divergence (HED):

$$D^{\mathrm{HE}}_{\alpha}(p:q) = D^{\mathrm{H}}_{\alpha,1,1}(p^E_{\alpha}:q^E_{\beta}) = D^{\mathrm{H}}_{\alpha,\frac{1}{\alpha},\frac{1}{\beta}}(p:q) = -\log\int_{\mathcal{X}} p(x)^{1/\alpha}q(x)^{1/\beta}\mathrm{d}x = B_{1/\alpha}(p:q), \tag{23}$$

which turns out to be the familiar skew Bhattacharyya divergence $B_{1/\alpha}(p:q)$; see [8].

Fact 5 (HED as a skew Bhattacharyya divergence)**.** *The Hölder escort divergence amounts to a skew Bhattacharyya divergence:* $D_\alpha^{\mathrm{HE}}(p:q) = B_{1/\alpha}(p:q)$ *for any* $\alpha > 0$.

In particular, the Cauchy–Schwarz escort divergence $\mathrm{CS}^{\mathrm{HE}}(p:q)$ amounts to the Bhattacharyya distance [25] $B(p:q) = -\log \int_{\mathcal{X}} \sqrt{p(x)q(x)}\mathrm{d}x$:

$$\mathrm{CS}^{\mathrm{HE}}(p:q) = D_2^{\mathrm{HE}}(p:q) = D_{2,1,1}^{\mathtt{H}}(p_2^E : q_2^E) = D_{2,\frac{1}{2},\frac{1}{2}}^{\mathtt{H}}(p:q) = B_{1/2}(p:q) = B(p:q). \tag{24}$$

Observe that the Cauchy–Schwarz escort distributions are the square root density representations [26] of distributions.

3. Proper Hölder Divergence

3.1. Definition

To get a proper HD between probability distributions $p(x)$ and $q(x)$, we need to have $p(x)^{\alpha\sigma} \propto q(x)^{\beta\tau}$. That is, we have $\alpha\sigma = \beta\tau$, or equivalently, we set $\tau = (\alpha-1)\sigma$ for free prescribed parameters $\alpha > 1$ and $\sigma > 0$. Alternatively, as we shall consider in the remainder, one may set $\alpha\sigma = \beta\tau = \gamma$ as a free prescribed parameter, which yields $\sigma = \gamma/\alpha$ and $\tau = \gamma/\beta$. Thus, in general, we define a bi-parametric family of proper Hölder divergence on probability distributions $D_{\alpha,\gamma}^{\mathtt{H}}$.

Let $p(x)$ and $q(x)$ be positive measures in $L^\gamma(\mathcal{X},\mu)$ for a prescribed scalar value $\gamma > 0$. Plugging $\sigma = \gamma/\alpha$ and $\tau = \gamma/\beta$ into the definition of HPD $D_{\alpha,\sigma,\tau}^{\mathtt{H}}$, we get the following definition:

Definition 2 (Proper Hölder divergence)**.** *For conjugate exponents* $\alpha, \beta > 0$ *and* $\gamma > 0$*, the proper Hölder divergence (HD) between two densities* $p(x)$ *and* $q(x)$ *is defined by:*

$$D_{\alpha,\gamma}^{\mathtt{H}}(p:q) = D_{\alpha,\frac{\gamma}{\alpha},\frac{\gamma}{\beta}}^{\mathtt{H}}(p:q) = -\log\left(\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta}\mathrm{d}x}{(\int_{\mathcal{X}} p(x)^\gamma \mathrm{d}x)^{1/\alpha}(\int_{\mathcal{X}} q(x)^\gamma \mathrm{d}x)^{1/\beta}}\right). \tag{25}$$

Following Hölder's inequality, we can check that $D_{\alpha,\gamma}^{\mathtt{H}}(p:q) \geq 0$ and $D_{\alpha,\gamma}^{\mathtt{H}}(p:q) = 0$ iff $p(x)^\gamma \propto q(x)^\gamma$, i.e., $p(x) \propto q(x)$ (see Appendix A). If $p(x)$ and $q(x)$ belong to the statistical probability manifold, then $D_{\alpha,\gamma}^{\mathtt{H}}(p:q) = 0$ iff $p(x) = q(x)$ almost everywhere. This says that HD is a proper divergence for probability measures, and it becomes a pseudo-divergence for positive measures. Note that we have abused the notation $D^{\mathtt{H}}$ to denote both the Hölder pseudo-divergence (with three subscripts) and the Hölder divergence (with two subscripts).

Similar to HPD, HD is asymmetric when $\alpha \neq \beta$ with the following reference duality:

$$D_{\alpha,\gamma}^{\mathtt{H}}(p:q) = D_{\bar{\alpha},\gamma}^{\mathtt{H}}(q:p). \tag{26}$$

HD can be symmetrized as:

$$S_{\alpha,\gamma}^{\mathtt{H}}(p:q) = \frac{D_{\alpha,\gamma}^{\mathtt{H}}(p:q) + D_{\alpha,\gamma}^{\mathtt{H}}(q:p)}{2} = -\frac{1}{2}\log\frac{\int_{\mathcal{X}} p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta}\mathrm{d}x \int_{\mathcal{X}} p(x)^{\gamma/\beta} q(x)^{\gamma/\alpha}\mathrm{d}x}{\int_{\mathcal{X}} p(x)^\gamma \mathrm{d}x \int_{\mathcal{X}} q(x)^\gamma \mathrm{d}x}. \tag{27}$$

Furthermore, one can easily check that HD is a projective divergence.

For conjugate exponents $\alpha, \beta > 0$ and $\gamma > 0$, we rewrite the definition of HD as:

$$\begin{aligned} D^{\mathrm{H}}_{\alpha,\gamma}(p:q) &= -\log \int_{\mathcal{X}} \left(\frac{p(x)^\gamma}{\int_{\mathcal{X}} p(x)^\gamma \mathrm{d}x} \right)^{1/\alpha} \left(\frac{q(x)^\gamma}{\int_{\mathcal{X}} q(x)^\gamma \mathrm{d}x} \right)^{1/\beta} \mathrm{d}x, \\ &= -\log \left(p^E_{1/\gamma}(x) \right)^{1/\alpha} \left(q^E_{1/\gamma}(x) \right)^{1/\beta} \mathrm{d}x = B_{\frac{1}{\alpha}}(p^E_{1/\gamma} : q^E_{1/\gamma}). \end{aligned}$$

Therefore, HD can be reinterpreted as the skew Bhattacharyya divergence [8] between the escort distributions. In particular, when $\gamma = 1$, we get:

$$D^{\mathrm{H}}_{\alpha,1}(p:q) = -\log \int_{\mathcal{X}} p(x)^{1/\alpha} q(x)^{1/\beta} \mathrm{d}x = B_{\frac{1}{\alpha}}(p:q). \tag{28}$$

Fact 6. *The bi-parametric family of statistical Hölder divergences $D^{\mathrm{H}}_{\alpha,\gamma}$ passes through the one-parametric family of skew Bhattacharyya divergences when $\gamma = 1$.*

3.2. Special Case: The Cauchy–Schwarz Divergence

Within the family of Hölder divergence, we set $\alpha = \beta = \gamma = 2$ and get the Cauchy–Schwarz (CS) divergence.

$$D^{\mathrm{H}}_{2,2}(p:q) = D^{\mathrm{H}}_{2,1,1}(p:q) = \mathrm{CS}(p:q). \tag{29}$$

Figure 1 displays a diagram of those divergence classes with their inclusion relationships.

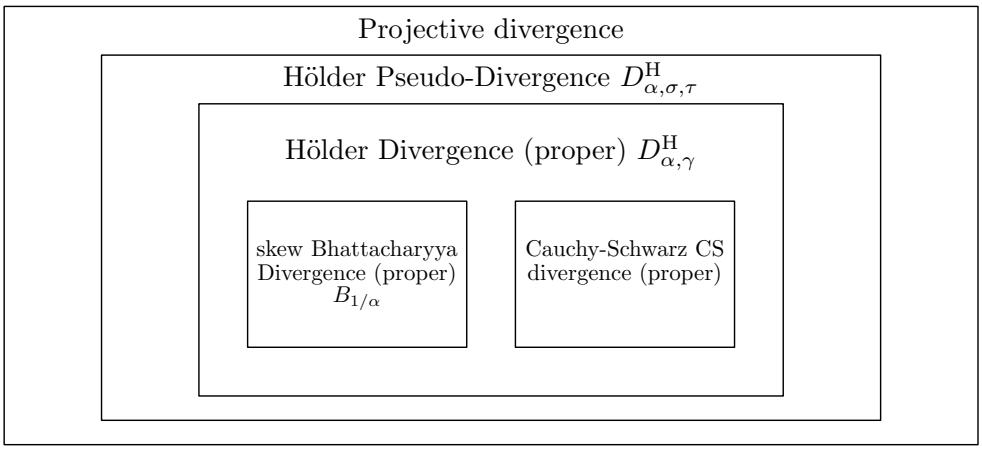


Figure 1. Hölder proper divergence (bi-parametric) and Hölder improper pseudo-divergence (tri-parametric) encompass Cauchy–Schwarz divergence and skew Bhattacharyya divergence.

As stated earlier, notice that the Cauchy–Schwarz inequality

$$\int p(x)q(x)\mathrm{d}x \leq \sqrt{\left(\int p(x)^2 \mathrm{d}x \right) \left(\int p(x)^2 \mathrm{d}x \right)} \tag{30}$$

is not proper as it is an equality when $p(x)$ and $q(x)$ are linearly dependent (i.e., $p(x) = \lambda q(x)$ for $\lambda > 0$). The arguments of the CS divergence are square-integrable real-valued density functions $p(x)$ and $q(x)$. Thus, the Cauchy–Schwarz divergence is not proper for positive measures, but is proper for normalized probability distributions, since in this case, $\int p(x)\mathrm{d}x = \int \lambda q(x)\mathrm{d}x = 1$ implies that $\lambda = 1$.

3.3. Limit Cases of Hölder Divergences and Statistical Estimation

Let us define the inner product of unnormalized densities as:

$$\langle \tilde{p}(x), \tilde{q}(x) \rangle = \int_{\mathcal{X}} \tilde{p}(x)\tilde{q}(x)\mathrm{d}x \tag{31}$$

(for $L^2(\mathcal{X},\mu)$ integrable functions), and define the L_α norm of densities as $\|\tilde{p}(x)\|_\alpha = (\int_{\mathcal{X}} \tilde{p}(x)^\alpha \mathrm{d}x)^{1/\alpha}$ for $\alpha \geq 1$. Then, the CS divergence can be concisely written as:

$$\mathrm{CS}(\tilde{p}:\tilde{q}) = -\log \frac{\langle \tilde{p}(x), \tilde{q}(x)\rangle}{\|\tilde{p}(x)\|_2 \|\tilde{q}(x)\|_2}, \tag{32}$$

and the Hölder pseudo-divergence is written as:

$$D^{\mathrm{H}}_{\alpha,1,1}(\tilde{p}:\tilde{q}) = -\log \frac{\langle \tilde{p}(x), \tilde{q}(x)\rangle}{\|\tilde{p}(x)\|_\alpha \|\tilde{q}(x)\|_{\bar{\alpha}}}. \tag{33}$$

When $\alpha \to 1^+$, we have $\bar{\alpha} = \alpha/(\alpha-1) \to +\infty$. Then, it comes that:

$$\lim_{\alpha\to 1^+} D^{\mathrm{H}}_{\alpha,1,1}(\tilde{p}:\tilde{q}) = -\log \frac{\langle \tilde{p}(x), \tilde{q}(x)\rangle}{\|\tilde{p}(x)\|_1 \|\tilde{q}(x)\|_\infty} = -\log \langle \tilde{p}(x), \tilde{q}(x)\rangle + \log \int_{\mathcal{X}} \tilde{p}(x)\mathrm{d}x + \log \max_{x\in\mathcal{X}} \tilde{q}(x). \tag{34}$$

When $\alpha \to +\infty$ and $\bar{\alpha} \to 1^+$, we have:

$$\begin{aligned} \lim_{\alpha\to+\infty} D^{\mathrm{H}}_{\alpha,1,1}(\tilde{p}:\tilde{q}) &= -\log \frac{\langle \tilde{p}(x), \tilde{q}(x)\rangle}{\|\tilde{p}(x)\|_\infty \|\tilde{q}(x)\|_1} \\ &= -\log \langle \tilde{p}(x), \tilde{q}(x)\rangle + \log \max_{x\in\mathcal{X}} \tilde{p}(x) + \log \int_{\mathcal{X}} \tilde{q}(x)\mathrm{d}x. \end{aligned} \tag{35}$$

Now, consider a pair of probability densities $p(x)$ and $q(x)$. We have:

$$\begin{aligned} \lim_{\alpha\to 1^+} D^{\mathrm{H}}_{\alpha,1,1}(p:q) &= -\log \langle p(x), q(x)\rangle + \max_{x\in\mathcal{X}} \log q(x), \\ \lim_{\alpha\to+\infty} D^{\mathrm{H}}_{\alpha,1,1}(p:q) &= -\log \langle p(x), q(x)\rangle + \max_{x\in\mathcal{X}} \log p(x), \\ \mathrm{CS}(p:q) &= -\log \langle p(x), q(x)\rangle + \log \|p(x)\|_2 + \log \|q(x)\|_2. \end{aligned} \tag{36}$$

In an estimation scenario, $p(x)$ is fixed, and $q(x\,|\,\theta) = q_\theta(x)$ is free along a parametric manifold $\mathcal{M}$; then, minimizing Hölder divergence reduces to:

$$\begin{aligned} \arg\min_{\theta\in\mathcal{M}} \lim_{\alpha\to 1^+} D^{\mathrm{H}}_{\alpha,1,1}(p:q_\theta) &= \arg\min_{\theta\in\mathcal{M}} \Big(-\log \langle p(x), q_\theta(x)\rangle + \max_{x\in\mathcal{X}} \log q_\theta(x) \Big), \\ \arg\min_{\theta\in\mathcal{M}} \lim_{\alpha\to+\infty} D^{\mathrm{H}}_{\alpha,1,1}(p:q) &= \arg\min_{\theta\in\mathcal{M}} \Big(-\log \langle p(x), q_\theta(x)\rangle \Big), \\ \arg\min_{\theta\in\mathcal{M}} \mathrm{CS}(p:q) &= \arg\min_{\theta\in\mathcal{M}} \Big(-\log \langle p(x), q_\theta(x)\rangle + \log \|q_\theta(x)\|_2 \Big). \end{aligned} \tag{37}$$

Therefore, when α varies from 1 to $+\infty$, only the regularizer in the minimization problem changes. In any case, Hölder divergence always has the term $-\log\langle p(x), q(x)\rangle$, which shares a similar form as the Bhattacharyya distance [25]:

$$B(p:q) = -\log \int_{\mathcal{X}} \sqrt{p(x)q(x)}\mathrm{d}x = -\log \langle \sqrt{p(x)}, \sqrt{q(x)}\rangle. \tag{38}$$

HPD between $\tilde{p}(x)$ and $\tilde{q}(x)$ is also closely related to their cosine similarity $\frac{\langle \tilde{p}(x), \tilde{q}(x)\rangle}{\|\tilde{p}(x)\|_2 \|\tilde{q}(x)\|_2}$. When $\alpha = 2$, $\sigma = \tau = 1$, HPD is exactly the cosine similarity after a non-linear transformation.

4. Closed-Form Expressions of HPD and HD for Conic and Affine Exponential Families

We report closed-form formulas for the HPD and HD between two distributions belonging to the same exponential family provided that the natural parameter space is a cone or affine. A cone Ω is a convex domain, such that for $P, Q \in \Omega$ and any $\lambda > 0$, we have $P + \lambda Q \in \Omega$. For example, the set of

positive measures absolutely continuous with a base measure μ is a cone. Recall that an exponential family [19] has a density function $p(x;\theta)$ that can be written canonically as:

$$p(x;\theta) = \exp\left(\langle t(x), \theta\rangle - F(\theta) + k(x)\right). \tag{39}$$

In this work, we consider the auxiliary carrier measure term $k(x) = 0$. The base measure is either the Lebesgue measure μ or the counting measure μ_C. A conic or affine exponential family (CAEF) is an exponential family with the natural parameter space Θ being a cone or affine. The log-normalizer $F(\theta)$ is a strictly convex function also called the cumulant generating function [3].

Lemma 1 (HPD and HD for CAEFs)**.** *For distributions $p(x;\theta_p)$ and $p(x;\theta_q)$ belonging to the same exponential family with conic or affine natural parameter space [23], both the HPD and HD are available in closed-form:*

$$D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) = \frac{1}{\alpha}F(\alpha\sigma\theta_p) + \frac{1}{\beta}F(\beta\tau\theta_q) - F(\sigma\theta_p + \tau\theta_q), \tag{40}$$

$$D^{\mathtt{H}}_{\alpha,\gamma}(p:q) = \frac{1}{\alpha}F(\gamma\theta_p) + \frac{1}{\beta}F(\gamma\theta_q) - F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right). \tag{41}$$

Proof. Consider $k(x) = 0$ and a conic or affine natural space Θ (see [23]); then, for all $a, b > 0$, we have:

$$\left(\int p(x)^a \mathrm{d}x\right)^{\frac{1}{b}} = \exp\left(\frac{1}{b}F(a\theta_p) - \frac{a}{b}F(\theta_p)\right), \tag{42}$$

since $a\theta_p \in \Theta$. Indeed, we have:

$$\begin{aligned}
\left(\int p(x)^a \mathrm{d}x\right)^{1/b} &= \left(\int \exp\left(\langle a\theta, t(x)\rangle - aF(\theta)\right)\mathrm{d}x\right)^{1/b} \\
&= \left(\int \exp\left(\langle a\theta, t(x)\rangle - F(a\theta) + F(a\theta) - aF(\theta)\right)\mathrm{d}x\right)^{1/b} \\
&= \exp\left(\frac{1}{b}F(a\theta) - \frac{a}{b}F(\theta)\right)\left(\underbrace{\int \exp\left(\langle a\theta, t(x)\rangle - F(a\theta)\right)\mathrm{d}x}_{=1}\right)^{1/b}.
\end{aligned}$$

Similarly, we have for all $a, b > 0$ (details omitted),

$$\int p(x)^a q(x)^b \mathrm{d}x = \exp(F(a\theta_p + b\theta_q) - aF(\theta_p) - bF(\theta_q)), \tag{43}$$

since $a\theta_p + b\theta_q \in \Theta$. Therefore, we get:

$$\begin{aligned}
D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) &= -\log\frac{\int p(x)^\sigma q(x)^\tau \mathrm{d}x}{\left(\int p(x)^{\alpha\sigma}\mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int q(x)^{\beta\tau}\mathrm{d}x\right)^{\frac{1}{\beta}}} \\
&= -F(\sigma\theta_p + \tau\theta_q) + F(\sigma\theta_p) + F(\tau\theta_q) + \frac{1}{\alpha}F(\alpha\sigma\theta_p) - F(\sigma\theta_p) + \frac{1}{\beta}F(\beta\tau\theta_q) - F(\tau\theta_q) \\
&= \frac{1}{\alpha}F(\alpha\sigma\theta_p) + \frac{1}{\beta}F(\beta\tau\theta_q) - F(\sigma\theta_p + \tau\theta_q) \geq 0, \\
D^{\mathtt{H}}_{\alpha,\gamma}(p:q) &= -\log\frac{\int p(x)^{\gamma/\alpha} q(x)^{\gamma/\beta}\mathrm{d}x}{\left(\int p(x)^\gamma \mathrm{d}x\right)^{\frac{1}{\alpha}}\left(\int q(x)^\gamma \mathrm{d}x\right)^{\frac{1}{\beta}}} \\
&= -F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right) + \frac{\gamma}{\alpha}F(\theta_p) + \frac{\gamma}{\beta}F(\theta_q) + \frac{1}{\alpha}F(\gamma\theta_p) - \frac{\gamma}{\alpha}F(\theta_p) + \frac{1}{\beta}F(\gamma\theta_q) - \frac{\gamma}{\beta}F(\theta_q) \\
&= \frac{1}{\alpha}F(\gamma\theta_p) + \frac{1}{\beta}F(\gamma\theta_q) - F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right) \geq 0.
\end{aligned}$$

When $1 > \alpha > 0$, we have $\beta = \frac{\alpha}{\alpha-1} < 0$. To get similar results for the reverse Hölder divergence, we need the natural parameter space to be affine (e.g., isotropic Gaussians or multinomials; see [27]). □

In particular, if $p(x)$ and $q(x)$ belong to the same exponential family so that $p(x) = \exp(\langle\theta_p, t(x)\rangle - F(\theta_p))$ and $q(x) = \exp(\langle\theta_q, t(x)\rangle - F(\theta_q))$, one can easily check that $D^{\mathrm{H}}_{\alpha,1,1}(p:q) = 0$ iff $\theta_q = (\alpha-1)\theta_p$. For HD, we can check that $D^{\mathrm{H}}_{\alpha,\gamma}(p:p) = 0$ is proper since $\frac{1}{\alpha} + \frac{1}{\beta} = 1$.

The following result is straightforward from Lemma 1.

Lemma 2 (Symmetric HPD and HD for CAEFs)**.** *For distributions $p(x;\theta_p)$ and $p(x;\theta_q)$ belonging to the same exponential family with conic or affine natural parameter space [23], the symmetric HPD and HD are available in closed-form:*

$$S^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) = \frac{1}{2}\left[\frac{1}{\alpha}F(\alpha\sigma\theta_p) + \frac{1}{\beta}F(\beta\tau\theta_p) + \frac{1}{\alpha}F(\alpha\sigma\theta_q) + \frac{1}{\beta}F(\beta\tau\theta_q) - F(\sigma\theta_p + \tau\theta_q) - F(\tau\theta_p + \sigma\theta_q)\right];$$
$$S^{\mathrm{H}}_{\alpha,\gamma}(p:q) = \frac{1}{2}\left[F(\gamma\theta_p) + F(\gamma\theta_q) - F\left(\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q\right) - F\left(\frac{\gamma}{\beta}\theta_p + \frac{\gamma}{\alpha}\theta_q\right)\right].$$

Remark 1. *By reference duality,*

$$S^{\mathrm{H}}_{\alpha,\sigma,\tau}(p:q) = S^{\mathrm{H}}_{\bar{\alpha},\tau,\sigma}(p:q);$$
$$S^{\mathrm{H}}_{\alpha,\gamma}(p:q) = S^{\mathrm{H}}_{\bar{\alpha},\gamma}(p:q).$$

Note that the Hölder score-induced divergence [18] does not admit in general closed-form formulas for exponential families since it relies on a function $\phi(\cdot)$ (see Definition 4 of [18]).

Note that CAEF convex log-normalizers satisfy:

$$\frac{1}{\alpha}F(\alpha\theta_p) + \frac{1}{\beta}F(\beta\theta_q) \geq F(\theta_p + \theta_q). \tag{44}$$

A necessary condition is that $F(\lambda\theta) \geq \lambda F(\theta)$ for $\lambda > 0$ (take $\theta_p = \theta$, $\theta_q = 0$ and $F(0) = 0$ in the above equality).

The escort distribution for an exponential family is given by:

$$p^E_\alpha(x;\theta) = e^{\frac{F(\theta)}{\alpha} - F(\frac{\theta}{\alpha})} p(x;\theta)^{\frac{1}{\alpha}}. \tag{45}$$

The Hölder equality holds when $p(x)^\alpha \propto q(x)^\beta$ or $p(x)^\alpha q(x)^{-\beta} \propto 1$. For exponential families, this condition is satisfied when $\alpha\theta_p - \beta\theta_q \in \Theta$. That is, we need to have:

$$\alpha\left(\theta_p - \frac{1}{\alpha-1}\theta_q\right) \in \Theta. \tag{46}$$

Thus, we may choose small enough $\alpha = 1 + \epsilon > 1$ so that the condition is not satisfied for fixed θ_p and θ_q for many exponential distributions. Since multinomials have affine natural space [27], this condition is always met, but not for non-affine natural parameter spaces like normal distributions.

Notice the following fact:

Fact 7 (Density of a CAEF in $L^\gamma(\mathcal{X},\mu)$)**.** *The density of exponential families with conic or affine natural parameter space belongs to $L^\gamma(\mathcal{X},\mu)$ for any $\gamma > 0$.*

Proof. We have $\int_{\mathcal{X}}(\exp(\langle\theta, t(x)\rangle - F(\theta)))^\gamma \mathrm{d}\mu(x) = e^{F(\gamma\theta) - \gamma F(\theta)} < \infty$ for any $\gamma > 0$ provided that $\gamma\theta$ belongs to the natural parameter space. When Θ is a cone or affine, the condition is satisfied. □

Let $\tilde{p}(x;\theta) = \exp\left(\langle t(x), \theta\rangle\right)$ denote the unnormalized positive exponential family density and $p(x;\theta) = \frac{\tilde{p}(x;\theta)}{Z(\theta)}$ the normalized density with $Z(\theta) = \exp(F(\theta))$ the partition function. Although HD is a projective divergence since we have $D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p_1 : p_2) = D^{\mathtt{H}}_{\alpha,\sigma,\tau}(\tilde{p}_1 : \tilde{p}_2)$, observe that the HD value depends on the log-normalizer $F(\theta)$ (since the HD is an integral on the support; see also [12] for a similar argument with the γ-divergence [11]).

In practice, even when the log-normalizer is computationally intractable, we may still estimate the HD by Monte Carlo techniques: Indeed, we can sample a distribution $\tilde{p}(x)$ either by rejection sampling [12] or by the Markov chain Monte Carlo (MCMC) Metropolis–Hasting technique: It just requires to be able to sample a proposal distribution that has the same support.

We shall now instantiate the HPD and HD formulas for several exponential families with conic or affine natural parameter spaces.

4.1. Case Study: Categorical Distributions

Let $p = (p_0, \cdots, p_m)$ and $q = (q_0, \cdots, q_m)$ be two categorical distributions in the m-dimensional probability simplex Δ^m. We rewrite p in the canonical form of exponential families [19] as:

$$p_i = \exp\left((\theta_p)_i - \log\left(1 + \sum_{i=1}^{m} \exp(\theta_p)_i\right)\right), \quad \forall i \in \{1, \cdots, m\}, \tag{47}$$

with the redundant parameter:

$$p_0 = 1 - \sum_{i=1}^{m} p_i = \frac{1}{1 + \sum_{i=1}^{m} \exp(\theta_p)_i}. \tag{48}$$

From Equation (47), the convex cumulant generating function has the form $F(\theta) = \log\left(1 + \sum_{i=1}^{m} \exp(\theta_p)_i\right)$. The inverse transformation from p to θ is therefore given by:

$$\theta_i = \log\left(\frac{p_i}{p_0}\right), \quad \forall i \in \{1, \cdots, m\}. \tag{49}$$

The natural parameter space Θ is affine (hence conic), and by applying Lemma 1, we get the following closed-form formula:

$$\begin{aligned} D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p : q) =& \frac{1}{\alpha} \log\left(1 + \sum_{i=1}^{m} \exp(\alpha\sigma(\theta_p)_i)\right) + \frac{1}{\beta} \log\left(1 + \sum_{i=1}^{m} \exp(\beta\tau(\theta_q)_i)\right) \\ & - \log\left(1 + \sum_{i=1}^{m} \exp(\sigma(\theta_p)_i + \tau(\theta_q)_i)\right), \end{aligned} \tag{50}$$

$$\begin{aligned} D^{\mathtt{H}}_{\alpha,\gamma}(p : q) =& \frac{1}{\alpha} \log\left(1 + \sum_{i=1}^{m} \exp(\gamma(\theta_p)_i)\right) + \frac{1}{\beta} \log\left(1 + \sum_{i=1}^{m} \exp(\gamma(\theta_q)_i)\right) \\ & - \log\left(1 + \sum_{i=1}^{m} \exp\left(\frac{\gamma}{\alpha}(\theta_p)_i + \frac{\gamma}{\beta}(\theta_q)_i\right)\right). \end{aligned} \tag{51}$$

To get some intuitions, Figure 2 shows the Hölder divergence from a given reference distribution p_r to any categorical distribution (p_0, p_1, p_2) in the 2D probability simplex Δ^2. A main observation is that the Kullback–Leibler (KL) divergence exhibits a barrier near the boundary $\partial\Delta^2$ with large values. This is not the case for Hölder divergences: $D^{\mathtt{H}}_{\alpha,1,1}(p_r : p)$ does not have a sharp increase near the boundary (although it still penalizes the corners of Δ^2). For example, let $p = (0, 1/2, 1/2)$, $p_r = (1/3, 1/3, 1/3)$, then $\mathrm{KL}(p_r : p) \to \infty$, but $D^{\mathtt{H}}_{2,1,1}(p_r : p) = \sqrt{2/3}$. Another observation is that the minimum $D(p_r : p)$ can be reached at some point $p \neq p_r$ (see for example $D^{\mathtt{H}}_{4,1,1}(p_r : p)$

in Figure 2b; the bluest area corresponding to the minimum of $D(p_r : p)$ is not in the same location as the reference point).

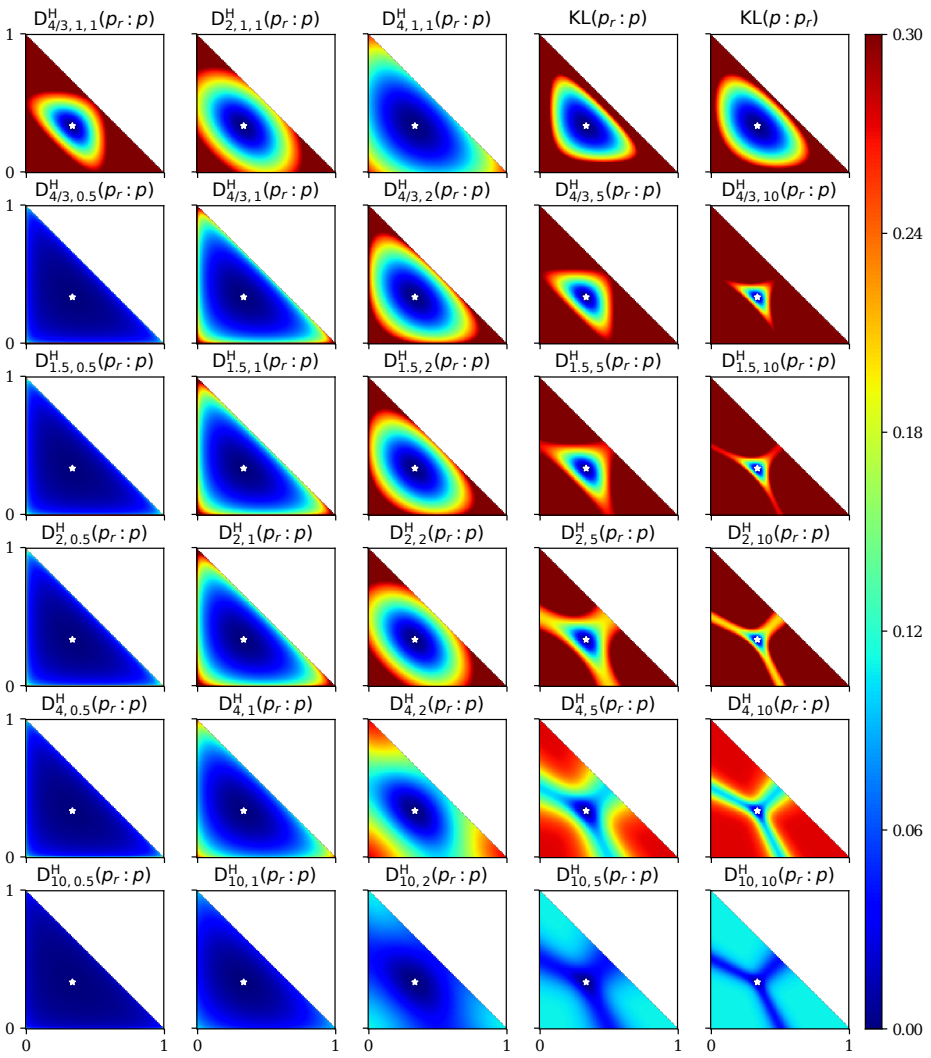
Consider an HPD ball of center c and prescribed radius r w.r.t. the HPD. Since $p(x)^{\alpha-1}$ for $\alpha \neq 2$ does not belong to the probability manifold, but to the positive measure manifold, and since the distance is projective, we deduce that the displaced ball center c' of a ball c lying in the probability manifold can be computed as the intersection of the ray $\lambda p(x)^{\alpha-1}$ anchored at the origin 0 and passing through $p(x)^{\alpha-1}$ with the probability manifold. For the discrete probability simplex Δ, since we have $\lambda \sum_{x\in\mathcal{X}} p(x)^{\alpha-1} = 1$, we deduce that the displaced ball center is at:

$$c' = \frac{c}{\sum_{x\in\mathcal{X}} p(x)^{\alpha-1}} \tag{52}$$

This center is displayed as "•" in Figure 2.

In general, the HPD bisector [28] between two distributions belonging to the same CAEF is defined by:

$$\frac{1}{\alpha}(F(\alpha\theta_1) - F(\alpha\theta_2)) = F(\theta_2 + \theta) - F(\theta_1 + \theta). \tag{53}$$



(**a**)

Figure 2. *Cont.*

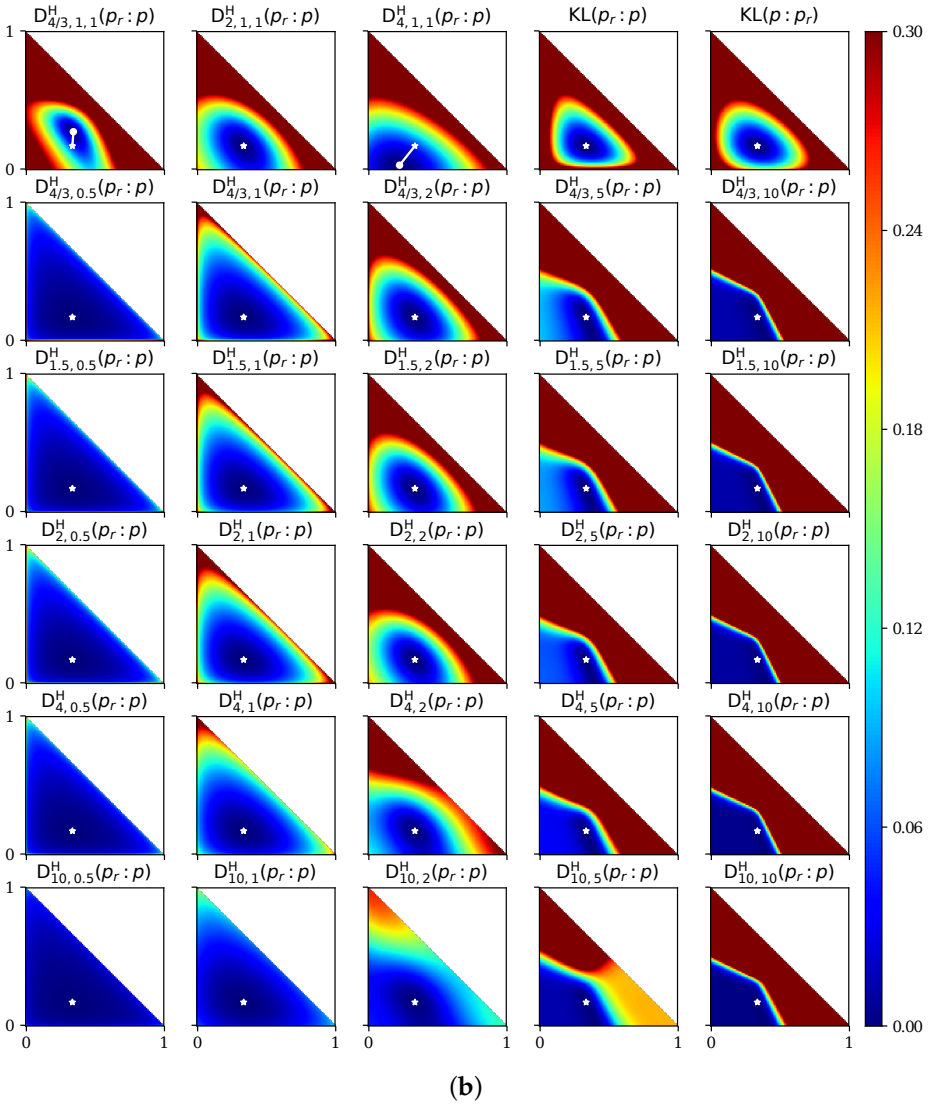


Figure 2. First row: the Hölder pseudo divergence (HPD) $D^{\mathrm{H}}_{\alpha,1,1}(p_r : p)$ for $\alpha \in \{4/3, 2, 4\}$, KL divergence and reverse KL divergence. Remaining rows: the HD $D^{\mathrm{H}}_{\alpha,\gamma}(p_r : p)$ for $\alpha \in \{4/3, 1.5, 2, 4, 10\}$ (from top to bottom) and $\gamma \in \{0.5, 1, 2, 5, 10\}$ (from left to right). The reference distribution p_r is presented as "$\star$". The minimizer of $D^{\mathrm{H}}_{\alpha,1,1}(p_r : p)$, if different from p_r, is presented as "$\bullet$". Notice that $D^{\mathrm{H}}_{2,2} = D^{\mathrm{H}}_{2,1,1}$. (**a**) Reference categorical distribution $p_r = (1/3, 1/3, 1/3)$; (**b**) reference categorical distribution $p_r = (1/2, 1/3, 1/6)$.

4.2. Case Study: Bernoulli Distribution

The Bernoulli distribution is just a special case of the category distribution when the number of categories is two (i.e., $m = 1$). To be consistent with the previous section, we rewrite a Bernoulli distribution $p = (p_0, p_1)$ in the canonical form:

$$p_1 = \exp\left(\theta_p - \log\left(1 + \exp(\theta_p)\right)\right) = \frac{\exp(\theta_p)}{1 + \exp(\theta_p)}, \tag{54}$$

so that:

$$p_0 = \frac{1}{1 + \exp(\theta_p)}. \tag{55}$$

Then, the cumulant generating function becomes $F(\theta_p) = \log\left(1+\exp(\theta_p)\right)$. By Lemma 1,

$$D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) = \frac{1}{\alpha}\log\left(1+\exp(\alpha\sigma\theta_p)\right) + \frac{1}{\beta}\log\left(1+\exp(\beta\tau\theta_q)\right) - \log\left(1+\exp(\sigma\theta_p+\tau\theta_q)\right), \quad (56)$$

$$D^{\mathtt{H}}_{\alpha,\gamma}(p:q) = \frac{1}{\alpha}\log\left(1+\exp(\gamma\theta_p)\right) + \frac{1}{\beta}\log\left(1+\exp(\gamma\theta_q)\right) - \log\left(1+\exp\left(\frac{\gamma}{\alpha}\theta_p+\frac{\gamma}{\beta}\theta_q\right)\right). \quad (57)$$

4.3. Case Study: MultiVariate Normal Distributions

Let us now instantiate the formulas for multivariate normals (Gaussian distributions). We have the log-normalizer $F(\theta)$ expressed using the usual parameters as [15]:

$$F(\theta) = F(\mu(\theta),\Sigma(\theta)) = \frac{1}{2}\log(2\pi)^d|\Sigma| + \frac{1}{2}\mu^\top\Sigma^{-1}\mu. \quad (58)$$

Since:

$$\theta = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1}) = (v, M), \quad \mu = -\frac{1}{2}M^{-1}v, \quad \Sigma = -\frac{1}{2}M^{-1}. \quad (59)$$

It follows that:

$$\theta_p+\theta_q = \theta_{p+q} = (v_p+v_q, M_p+M_q) = \left(\Sigma_p^{-1}\mu_p+\Sigma_q^{-1}\mu_q, -\frac{1}{2}\Sigma_p^{-1}-\frac{1}{2}\Sigma_q^{-1}\right). \quad (60)$$

Therefore, we have:

$$\mu_{p+q} = (\Sigma_p^{-1}+\Sigma_q^{-1})^{-1}(\Sigma_p^{-1}\mu_p+\Sigma_q^{-1}\mu_q), \quad \Sigma_{p+q} = (\Sigma_p^{-1}+\Sigma_q^{-1})^{-1} \quad (61)$$

We thus get the following closed-form formula for $p \sim N(\mu_p,\Sigma_p)$ and $q \sim N(\mu_q,\Sigma_q)$:

$$\begin{aligned}
D^{\mathtt{H}}_{\alpha,\sigma,\tau}(N(\mu_p,\Sigma_p):N(\mu_q,\Sigma_q)) &= \frac{1}{2\alpha}\log\left|\frac{\Sigma_p}{\alpha\sigma}\right| + \frac{\sigma}{2}\mu_p^\top\Sigma_p^{-1}\mu_p + \frac{1}{2\beta}\log\left|\frac{\Sigma_q}{\beta\tau}\right| + \frac{\tau}{2}\mu_q^\top\Sigma_q^{-1}\mu_q \\
&+ \frac{1}{2}\log\left|\sigma\Sigma_p^{-1}+\tau\Sigma_q^{-1}\right| - \frac{1}{2}(\sigma\Sigma_p^{-1}\mu_p+\tau\Sigma_q^{-1}\mu_q)^\top(\sigma\Sigma_p^{-1}+\tau\Sigma_q^{-1})^{-1}(\sigma\Sigma_p^{-1}\mu_p+\tau\Sigma_q^{-1}\mu_q); \\
D^{\mathtt{H}}_{\alpha,\gamma}(N(\mu_p,\Sigma_p):N(\mu_q,\Sigma_q)) &= \frac{1}{2\alpha}\log\left|\frac{\Sigma_p}{\gamma}\right| + \frac{\gamma}{2\alpha}\mu_p^\top\Sigma_p^{-1}\mu_p + \frac{1}{2\beta}\log\left|\frac{\Sigma_q}{\gamma}\right| + \frac{\gamma}{2\beta}\mu_q^\top\Sigma_q^{-1}\mu_q \\
&+ \frac{1}{2}\log\left|\frac{\gamma}{\alpha}\Sigma_p^{-1}+\frac{\gamma}{\beta}\Sigma_q^{-1}\right| - \frac{1}{2}\left(\frac{\gamma}{\alpha}\Sigma_p^{-1}\mu_p+\frac{\gamma}{\beta}\Sigma_q^{-1}\mu_q\right)^\top\left(\frac{\gamma}{\alpha}\Sigma_p^{-1}+\frac{\gamma}{\beta}\Sigma_q^{-1}\right)^{-1}\left(\frac{\gamma}{\alpha}\Sigma_p^{-1}\mu_p+\frac{\gamma}{\beta}\Sigma_q^{-1}\mu_q\right).
\end{aligned}$$

Figure 3 shows HPD and HD for univariate Gaussian distributions as compared to the KL divergence. Again, HPD and HD have more tolerance for distributions near the boundary $\sigma = 0$, which is in contrast to the (reverse) KL divergence.

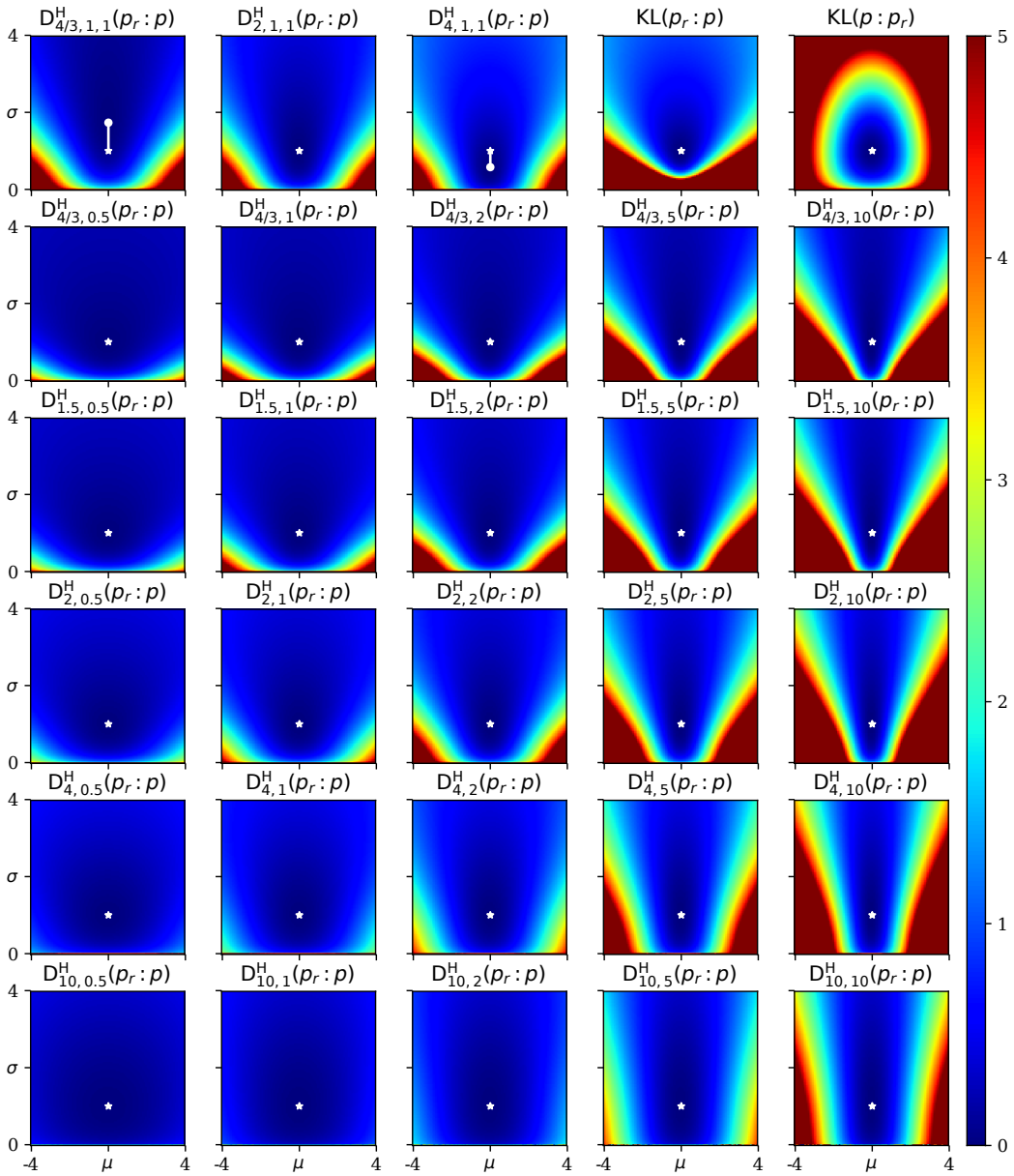


Figure 3. First row: $D^{\mathrm{H}}_{\alpha,1,1}(p_r:p)$, where p_r is the standard Gaussian distribution and $\alpha \in \{4/3, 2, 4\}$ compared to the KL divergence. The rest of the rows: $D^{\mathrm{H}}_{\alpha,\gamma}(p_r:p)$ for $\alpha \in \{4/3, 1.5, 2, 4, 10\}$ (from top to bottom) and $\gamma \in \{0.5, 1, 2, 5, 10\}$ (from left to right). Notice that $D^{\mathrm{H}}_{2,2} = D^{\mathrm{H}}_{2,1,1}$. The coordinate system is formed by μ (mean) and σ (standard deviation).

4.4. Case Study: Zero-Centered Laplace Distribution

The zero-centered Laplace distribution is defined on the support $(-\infty, \infty)$ with the pdf:

$$p(x;s) = \frac{1}{2s}\exp\left(-\frac{|x|}{s}\right) = \exp\left(-\frac{|x|}{s} - \log(2s)\right). \tag{62}$$

We have $\theta = -\frac{1}{s}$, $F(\theta) = \log(-\frac{2}{\theta})$. Therefore, it comes that:

$$\begin{aligned} D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q) &= \frac{1}{\alpha}\log\left(-\frac{2}{\alpha\sigma\theta_p}\right) + \frac{1}{\beta}\log\left(-\frac{2}{\beta\tau\theta_q}\right) - \log\left(-\frac{2}{\sigma\theta_p + \tau\theta_q}\right) \\ &= \frac{1}{\alpha}\log\left(\frac{s_p}{\alpha\sigma}\right) + \frac{1}{\beta}\log\left(\frac{s_q}{\beta\tau}\right) + \log\left(\frac{\sigma}{s_p} + \frac{\tau}{s_q}\right), \end{aligned} \tag{63}$$

$$\begin{aligned} D^{\mathtt{H}}_{\alpha,\gamma}(p:q) &= \frac{1}{\alpha}\log\left(-\frac{2}{\gamma\theta_p}\right) + \frac{1}{\beta}\log\left(-\frac{2}{\gamma\theta_q}\right) - \log\left(-\frac{2}{\frac{\gamma}{\alpha}\theta_p + \frac{\gamma}{\beta}\theta_q}\right) \\ &= \frac{1}{\alpha}\log s_p + \frac{1}{\beta}\log s_q + \log\left(\frac{1}{\alpha s_p} + \frac{1}{\beta s_p}\right). \end{aligned} \tag{64}$$

In this special case, $D^{\mathtt{H}}_{\alpha,\gamma}(p:q)$ does not vary with γ.

4.5. Case Study: Wishart Distribution

The Wishart distribution is defined on the $d \times d$ positive definite cone with the density:

$$p(X;n,S) = \frac{|X|^{\frac{n-d-1}{2}}\exp\left(-\frac{1}{2}\mathrm{tr}(S^{-1}X)\right)}{2^{\frac{nd}{2}}|S|^{\frac{n}{2}}\Gamma_d\left(\frac{n}{2}\right)}, \tag{65}$$

where $n > d-1$ is the degree of freedom and $S \succ 0$ is a positive-definite scale matrix. We rewrite it in the canonical form:

$$p(X;n,S) = \exp\left(-\frac{1}{2}\mathrm{tr}(S^{-1}X) + \frac{n-d-1}{2}\log|X| - \frac{nd}{2}\log 2 - \frac{n}{2}\log|S| - \log\Gamma_d\left(\frac{n}{2}\right)\right). \tag{66}$$

We can see that $\theta = (\theta^1, \theta^2)$, $\theta^1 = -\frac{1}{2}S^{-1}$, $\theta^2 = \frac{n-d-1}{2}$, and:

$$\begin{aligned} F(\theta) &= \frac{nd}{2}\log 2 + \frac{n}{2}\log|S| + \log\Gamma_d\left(\frac{n}{2}\right) \\ &= (\theta^2 + \frac{d+1}{2})d\log 2 + (\theta^2 + \frac{d+1}{2})\log\left|-\frac{1}{2}(\theta^1)^{-1}\right| + \log\Gamma_d\left(\theta^2 + \frac{d+1}{2}\right). \end{aligned} \tag{67}$$

The resulting $D^{\mathtt{H}}_{\alpha,\sigma,\tau}(p:q)$ and $D^{\mathtt{H}}_{\alpha,\gamma}(p:q)$ are straightforward from the above expression of $F(\theta)$ and Lemma 1. We will omit these tedious expressions for brevity.

4.6. Approximating Hölder Projective Divergences for Statistical Mixtures

Given two finite mixture models $m(x) = \sum_{i=1}^{k} w_i p_i(x)$ and $m'(x) = \sum_{j=1}^{k'} w'_j p'_j(x)$, we derive analytic bounds of their Hölder divergences. When only an approximation is needed, one may compute Hölder divergences based on Monte Carlo stochastic sampling.

Let us assume that all mixture components are in an exponential family [19], so that $p_i(x) = p(x;\theta_i) = \exp(\langle\theta_i, t(x)\rangle - F(\theta_i))$ and $p'_j(x) = p(x;\theta'_j) = \exp(\langle\theta'_j, t(x)\rangle - F(\theta'_j))$ are densities (w.r.t. the Lebesgue measure μ).

Without loss of generality, we only consider the pseudo Hölder divergence $D^{\mathtt{H}}_{\alpha,1,1}$. We rewrite it in the form:

$$D^{\mathtt{H}}_{\alpha,1,1}(m:m') = -\log\int_{\mathcal{X}} m(x)m'(x)\mathrm{d}x + \frac{1}{\alpha}\log\int_{\mathcal{X}} m(x)^{\alpha}\mathrm{d}x + \frac{1}{\beta}\log\int_{\mathcal{X}} m'(x)^{\beta}\mathrm{d}x. \tag{68}$$

To compute the first term, we observe that a product of mixtures is also a mixture:

$$\begin{aligned}\int_{\mathcal{X}} m(x)m'(x)\mathrm{d}x &= \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j \int_{\mathcal{X}} p_i(x)p'_j(x)\mathrm{d}x \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j \int_{\mathcal{X}} \exp\left(\langle \theta_i + \theta'_j, t(x)\rangle - F(\theta_i) - F(\theta'_j)\right)\mathrm{d}x \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j \exp\left(F(\theta_i+\theta'_j) - F(\theta_i) - F(\theta'_j)\right), \end{aligned} \tag{69}$$

which can be computed in $O(kk')$ time.

The second and third terms in Equation (68) are not straightforward to calculate and shall be bounded. Based on computational geometry, we adopt the log-sum-exp bounding technique of [29] and divide the support $\mathcal{X}$ into L pieces of elementary intervals $\mathcal{X} = \uplus_{l=1}^{L} I_l$. In each interval I_l, the indices:

$$\delta_l = \arg\max_i w_i p_i(x) \text{ and } \epsilon_l = \arg\min_i w_i p_i(x) \tag{70}$$

represent the unique dominating component and the dominated component. Then, we bound as follows:

$$\max\left\{\int_{I_l} k^\alpha w_{\epsilon_l}^\alpha p_{\epsilon_l}(x)^\alpha \mathrm{d}x, \int_{I_l} w_{\delta_l}^\alpha p_{\delta_l}(x)^\alpha \mathrm{d}x\right\} \leq \int_{I_l} m(x)^\alpha \mathrm{d}x \leq \int_{I_l} k^\alpha w_{\delta_l}^\alpha p_{\delta_l}(x)^\alpha \mathrm{d}x. \tag{71}$$

All terms on the lhs and rhs of Equation (71) can be computed exactly by noticing that:

$$\int_I p_i(x)^\alpha dx = \int_I \exp(\langle \alpha\theta_i, t(x)\rangle - \alpha F(\theta_i)) = \exp(F(\alpha\theta_i) - \alpha F(\theta_i)) \int_I p(x;\alpha\theta_i)\mathrm{d}x. \tag{72}$$

When $\alpha\theta \in \Theta$ where Θ denotes the natural parameter space, the integral $\int_I p(x;\alpha\theta_i)\mathrm{d}x$ converges; see [29] for further details.

Then, the bounds of $\int_{\mathcal{X}} m(x)^\alpha \mathrm{d}x$ can be obtained by summing the bounds in Equation (71) over all elementary intervals. Thus, $D^{\mathrm{H}}_{\alpha,1,1}(m:m')$ can be both lower and upper bounded.

5. Hölder Centroids and Center-Based Clustering

We study the application of HPD and HD for clustering distributions [30], specially clustering Gaussian distributions [31–33], which have been used in sound processing [31], sensor network [32], statistical debugging [32], quadratic invariants of switched systems [34], etc. Other potential applications of HD may include nonnegative matrix factorization [35], and clustering von Mises–Fisher [36,37] (log-normalizer expressed using Bessel functions).

5.1. Hölder Centroids

We study center-based clustering of a finite set of distributions belonging to the same exponential family. By a slight abuse of notation, we shall write $D^{\mathrm{H}}_{\alpha,\sigma,\tau}(\theta:\theta')$ instead of $D^{\mathrm{H}}_{\alpha,\sigma,\tau}(p_\theta : p_{\theta'})$. Given a list of distributions belonging to the same conic exponential family with natural parameters $\{\theta_1,\cdots,\theta_n\}$ and their associated positive weights $\{w_1,\cdots,w_n\}$ with $\sum_{i=1}^n w_i = 1$, consider their centroids based on HPD and HD as follows:

$$C_\alpha(\{\theta_i,w_i\}) = \arg\min_C \sum_{i=1}^n w_i D^{\mathrm{H}}_{\alpha,1,1}(\theta_i : C), \tag{73}$$

$$C_{\alpha,\gamma}(\{\theta_i,w_i\}) = \arg\min_C \sum_{i=1}^n w_i D^{\mathrm{H}}_{\alpha,\gamma}(\theta_i : C). \tag{74}$$

By an abuse of notation, C denotes both the HPD centroid and HD centroid. When the context is clear, the parameters in parentheses can be omitted so that these centroids are simply denoted as C_α and $C_{\alpha,\gamma}$. Both of them are defined as the right-sided centroids. The corresponding left-handed centroids are obtained according to reference duality, i.e.,

$$C_{\bar{\alpha}} = \arg\min_C \sum_{i=1}^n w_i D^{\mathtt{H}}_{\alpha,1,1}(C : \theta_i), \tag{75}$$

$$C_{\bar{\alpha},\gamma} = \arg\min_C \sum_{i=1}^n w_i D^{\mathtt{H}}_{\alpha,\gamma}(C : \theta_i). \tag{76}$$

By Lemma 1, these centroids can be obtained for distributions belonging to the same exponential family as follows:

$$C_\alpha = \arg\min_C \left[\frac{1}{\beta} F(\beta C) - \sum_{i=1}^n w_i F(\theta_i + C) \right], \tag{77}$$

$$C_{\alpha,\gamma} = \arg\min_C \left[\frac{1}{\beta} F(\gamma C) - \sum_{i=1}^n w_i F\left(\frac{\gamma}{\alpha}\theta_i + \frac{\gamma}{\beta} C \right) \right]. \tag{78}$$

Let $\gamma = \alpha$; we get:

$$C_{\alpha,\alpha}(\{\theta_i, w_i\}) = \arg\min_C \left[\frac{1}{\beta} F(\alpha C) - \sum_{i=1}^n w_i F\left(\theta_i + \frac{\alpha}{\beta} C \right) \right] = \frac{\beta}{\alpha} C_\alpha = \frac{1}{\alpha - 1} C_\alpha(\{\theta_i, w_i\}), \tag{79}$$

meaning that the HPD centroid is just a special case of HD centroid up to a scaling transformation in the natural parameters space. Let $\gamma = \beta$; we get:

$$C_{\alpha,\beta}(\{\theta_i, w_i\}) = \arg\min_C \left[\tfrac{1}{\beta} F(\beta C) - \textstyle\sum_{i=1}^n w_i F\left(\frac{\beta}{\alpha}\theta_i + C \right) \right] = C_\alpha\left(\left\{ \tfrac{\beta}{\alpha}\theta_i, w_i \right\} \right) = C_\alpha\left(\left\{ \tfrac{1}{\alpha-1}\theta_i, w_i \right\} \right). \tag{80}$$

Let us consider the general HD centroid $C_{\alpha,\gamma}$. Since F is convex, the minimization energy is the sum of a convex function $\frac{1}{\beta}F(\gamma C)$ with a concave function $-\sum_{i=1}^n w_i F\left(\frac{\gamma}{\alpha}\theta_i + \frac{\gamma}{\beta} C \right)$. We can therefore use the concave-convex procedure (CCCP) [8] that optimizes the difference of convex programs (DCPs): We start with $C^0_{\alpha,\gamma} = \sum_{i=1}^n w_i \theta_i$ (the barycenter, belonging to Θ) and then update:

$$C^{t+1}_{\alpha,\gamma} = \frac{1}{\gamma} (\nabla F)^{-1} \left(\sum_{i=1}^n w_i \nabla F\left(\frac{\gamma}{\alpha}\theta_i + \frac{\gamma}{\beta} C^t_{\alpha,\gamma} \right) \right) \tag{81}$$

for $t = 0, 1, \cdots$ until convergence. This can be done by noting that $\eta = \nabla F(\theta)$ are the dual parameters that are also known as the expectation parameters (or moment parameters). Therefore, ∇F and $(\nabla F)^{-1}$ can be computed through Legendre transformations between the natural parameter space and the dual parameter space.

This iterative optimization is guaranteed to converge to a local minimum, with a main advantage of bypassing the learning rate parameter of gradient descent algorithms. Since F is strictly convex, ∇F is monotonous, and the rhs expression can be interpreted as a multi-dimensional quasi-arithmetic mean. In fact, it is a barycenter on unnormalized weights scaled by $\beta = \bar{\alpha}$.

For exponential families, the symmetric HPD centroid is:

$$O_\alpha = \arg\min_O \sum_{i=1}^n w_i S^{\mathtt{H}}_{\alpha,1,1}(\theta_i : O) = \arg\min_O \left[\frac{1}{2\alpha} F(\alpha O) + \frac{1}{2\beta} F(\beta O) - \sum_{i=1}^n w_i F(\theta_i + O) \right]. \tag{82}$$

In this case, the CCCP update rule is not in closed form because we cannot easily inverse the sum of gradients (but when $\alpha = \beta$, the two terms collapse, so the CS centroid can be calculated using CCCP).

Nevertheless, we can implement the reciprocal operation numerically. Interestingly, the symmetric HD centroid can be solved by CCCP! It amounts to solving:

$$\begin{aligned} O_{\alpha,\gamma} &= \arg\min_{O} \sum_{i=1}^{n} w_i S^{\mathtt{H}}_{\alpha,\gamma}(\theta_i : O) \\ &= \arg\min_{O} \left[F(\gamma O) - \sum_{i=1}^{n} w_i \left(F\left(\frac{\gamma}{\alpha}\theta_i + \frac{\gamma}{\beta}O\right) + F\left(\frac{\gamma}{\beta}\theta_i + \frac{\gamma}{\alpha}O\right)\right)\right]. \end{aligned} \tag{83}$$

One can apply CCCP to iteratively update the centroid based on:

$$O^{t+1}_{\alpha,\gamma} = \frac{1}{\gamma}(\nabla F)^{-1}\left[\sum_{i=1}^{n} w_i \left(\frac{1}{\beta}\nabla F\left(\frac{\gamma}{\alpha}\theta_i + \frac{\gamma}{\beta}O^t_{\alpha,\gamma}\right) + \frac{1}{\alpha}\nabla F\left(\frac{\gamma}{\beta}\theta_i + \frac{\gamma}{\alpha}O^t_{\alpha,\gamma}\right)\right)\right]. \tag{84}$$

Notice the similarity with the updating procedure of $C^t_{\alpha,\gamma}$.
Once the centroid, say $O_{\alpha,\gamma}$, has been computed, we calculate the associated Hölder information:

$$\sum_{i=1}^{n} w_i S^{\mathtt{H}}_{\alpha,\gamma}(\theta_i : O_{\alpha,\gamma}), \tag{85}$$

which generalizes the notion of variance and Bregman information [5] to the case of Hölder distances.

5.2. Clustering Based on Symmetric Hölder Divergences

Given a set of fixed densities $\{p_1, \cdots, p_n\}$, we can perform variational k-means [6] with respect to the Hölder divergence to minimize the cost function:

$$E(O_1, \cdots, O_L, l_1, \cdots, l_n) = \sum_{i=1}^{n} S^{\mathtt{H}}_{\alpha,\gamma}(p_i : O_{l_i}), \tag{86}$$

where $O_1, \cdots, O_L$ are the cluster centers and $l_i \in \{1, \cdots, L\}$ is the cluster label of p_i. The algorithm is given by Algorithm 1. Notice that one does not need to wait for the CCCP iterations to converge. It only has to improve the cost function E before updating the assignment. We have implemented the algorithm based on the symmetric HD. One can easily modify it based on HPD and other variants.

Algorithm 1: Hölder variational k-means.

Input: A list of probability distributions $p_1, \cdots, p_n$; number of clusters L; $\alpha > 1$; $\gamma > 0$
Output: A clustering scheme $p_i \to \{1, \cdots, L\}, \forall i \in \{1, \cdots, n\}$

Randomly pick L distributions as the cluster centers $\{O_l\}_{l=1}^{L}$
while *not converged* **do**
  for $i = 1, \ldots, n$ **do**
    Assign $l_i = \arg\min_l S^{\mathtt{H}}_{\alpha,\gamma}(p_i : O_l)$
  for $l = 1, \ldots, L$ **do**
    `/* Variational k-means: Carry CCCP iterations until the current center improves the former cluster Hölder information */`
    Compute the centroid $O_l = \arg\min_O \sum_{i:l_i=l} S^{\mathtt{H}}_{\alpha,\gamma}(p_i : O)$
return $\{l_i\}_{i=1}^{n}$

We made a toy dataset generator, which can randomly generate n 2D Gaussians that have an underlying structure of two or three clusters. In the first cluster, the mean of each Gaussian $G(\mu, \Sigma)$ has the prior distribution $\mu \sim G((-2, 0), I)$; the covariance matrix is obtained by first generating $\sigma_1 \sim \Gamma(7, 0.01)$, $\sigma_2 \sim \Gamma(7, 0.003)$, where Γ means a gamma distribution with prescribed shape and scale,

then rotating the covariance matrix $\mathrm{diag}(\sigma_1, \sigma_2)$ so that the resulting Gaussian has a "radial direction" with respect to the center $(-2, 0)$. The second and third clusters are similar to the first cluster with the only difference being that their μ's are centered around $(2, 0)$ and $(0, 2\sqrt{3})$, respectively. See Figure 4 for an intuitive presentation of the toy dataset.

To reduce the number of parameters that has to be tuned, we only investigate the case $\alpha = \gamma$. If we choose $\alpha = \gamma = 2$, then $S^{\mathrm{H}}_{\alpha,\gamma}$ becomes the CS divergence, and Algorithm 1 reduces to traditional CS clustering. From Figure 4, we can observe that the clustering result does vary with the settings of α and γ. We performed clustering experiments on two different settings of the number of clusters and two different settings of the sample size. Table 1 shows the clustering accuracy measured by the percentage of "correctly-clustered" Gaussians, i.e., the output label by clustering algorithms that coincides with the true label corresponding to the data generating process. The large variance of the clustering accuracy is because different runs are based on different random datasets using the same generator. We see that the symmetric Hölder divergence can give significantly better clustering results as compared to CS clustering. Intuitively, the symmetric Hölder centroid with α and γ close to one has a smaller variance (see Figure 4); therefore, it can better capture the clustering structure. This hints that one should consider the general Hölder divergence to replace CS in similar clustering applications [22,38]. Although one faces the problem of tuning the parameter α and γ, Hölder divergences can potentially give better results. This is expected because CS is just one particular case of the class of Hölder divergences.

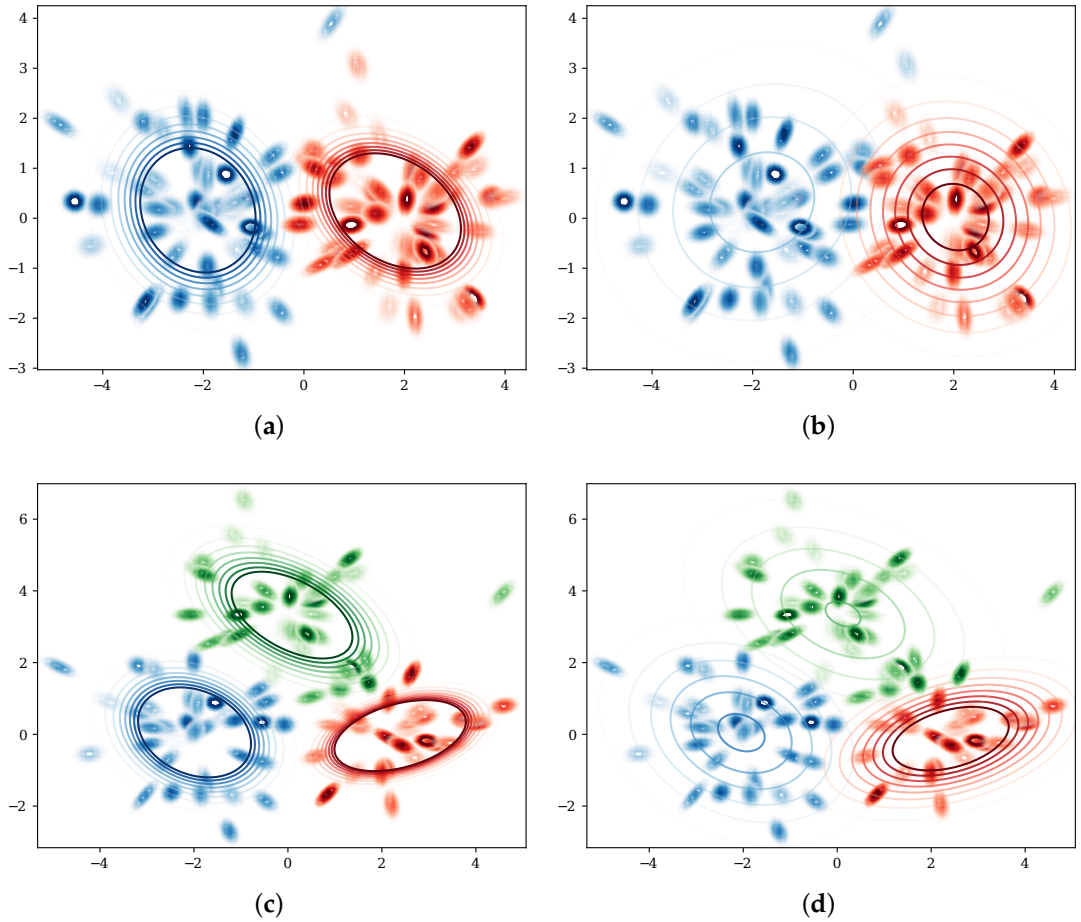


Figure 4. Variational k-means clustering results on a toy dataset consisting of a set of 2D Gaussians organized into two or three clusters. The cluster centroids are represented by contour plots using the same density levels. (**a**) $\alpha = \gamma = 1.1$ (Hölder clustering); (**b**) $\alpha = \gamma = 2$ (Cauchy–Schwarz clustering); (**c**) $\alpha = \gamma = 1.1$ (Hölder clustering); (**d**) $\alpha = \gamma = 2$ (Cauchy–Schwarz clustering).

Table 1. Clustering accuracy of the 2D Gaussian dataset (based on 1000 independent runs). CS, Cauchy–Schwarz. Bold numbers indicate the best obtained performance.

k (#Clusters)	n (#Samples)	$\alpha = \gamma = 1.1$	$\alpha = \gamma = 1.5$	$\alpha = \gamma = 2$ (CS)	$\alpha = \gamma = 10$
2	50	$\mathbf{94.5\% \pm 10.5\%}$	$89.9\% \pm 13.2\%$	$89.4\% \pm 13.5\%$	$88.9\% \pm 14.0\%$
	100	$\mathbf{96.9\% \pm 6.8\%}$	$94.3\% \pm 9.9\%$	$93.8\% \pm 10.6\%$	$93.1\% \pm 11.6\%$
3	50	$\mathbf{84.6\% \pm 15.5\%}$	$79.3\% \pm 14.8\%$	$79.0\% \pm 14.7\%$	$78.7\% \pm 14.5\%$
	100	$\mathbf{89.6\% \pm 13.8\%}$	$83.9\% \pm 14.6\%$	$83.1\% \pm 14.5\%$	$82.8\% \pm 14.4\%$

6. Conclusions and Perspectives

We introduced the notion of pseudo-divergences that generalizes the concept of divergences in information geometry [3] that are smooth non-metric statistical distances that are not required to obey the law of the indiscernibles. Pseudo-divergences can be built from inequalities by considering the inequality difference gap or its log-ratio gap. We then defined two classes of statistical measures based on Hölder's ordinary and reverse inequalities: the tri-parametric family of Hölder pseudo-divergences and the bi-parametric family of Hölder divergences. By construction, the Hölder divergences are proper divergences between probability densities. Both statistical Hölder distance families are projective divergences that do not require distributions to be normalized and admit closed-form expressions when considering exponential families with conic or affine natural parameter space (like multinomials or multivariate normals). Those two families of distances can be symmetrized and encompass both the Cauchy–Schwarz divergence and the family of skew Bhattacharyya divergences. Since the Cauchy–Schwarz divergence is often used in distribution clustering applications [22], we carried out preliminary experiments demonstrating experimentally that the symmetrized Hölder divergences improved over the Cauchy–Schwarz divergence for a toy dataset of Gaussians. We briefly touched upon the use of these novel divergences in statistical estimation theory. These projective Hölder (pseudo-)divergences are different from the recently introduced compositive scored-induced Hölder divergences [17,18] that are not projective divergences and do not admit closed-form expressions for exponential families in general.

We elicited the special role of escort distributions [3] for Hölder divergences in our framework: Escort distributions transform distributions to allow one:

- To reveal that Hölder pseudo-divergences on escort distributions amount to skew Bhattacharyya divergences [8],
- To transform the improper Hölder pseudo-divergences into proper Hölder divergences, and vice versa.

It is interesting to consider other potential applications of Hölder divergences and compare their efficiency against the reference Cauchy–Schwarz divergence: For example, HD t-SNE (Stochastic Neighbor Embedding) compared to CS t-SNE [39], HD vector quantization (VQ) compared to CS VQ [40], HD saliency vs. CS saliency detection in images [41], etc.

Let us conclude with a perspective note on pseudo-divergences, statistical estimators and manifold learning. Proper divergences have been widely used in statistical estimators to build families of estimators [42,43]. Similarly, given a prescribed density $p_0(x)$, a pseudo-divergence yields a corresponding estimator by minimizing $D(p_0 : q)$ with respect to $q(x)$. However, in this case, the resulting $q(x)$ is potentially biased and is not guaranteed to recover the optimal input $p_0(x)$. Furthermore, the minimizer of $D(p_0 : q)$ may not be unique, i.e., there could be more than one probability density $q(x)$ yielding $D(p_0 : q) = 0$.

How can pseudo-divergences be useful? We have the following two simple arguments:

- In an estimation scenario, we can usually pre-compute $p_1(x) \neq p_0(x)$ according to $D(p_1 : p_0) = 0$. Then, the estimation $q(x) = \arg\min_q D(p_1 : q)$ will automatically target at $p_0(x)$. We call this technique "pre-aim."

For example, given positive measure $p(x)$, we first find $p_0(x)$ to satisfy $D^{\mathtt{H}}_{\alpha,1,1}(p_0 : p) = 0$. We have $p_0(x) = p(x)^{\frac{1}{\alpha-1}}$ that satisfies this condition. Then, a proper divergence between $p(x)$ and $q(x)$ can be obtained by aiming $q(x)$ towards $p_0(x)$. For conjugate exponents α and β,

$$\begin{aligned} D^{\mathtt{H}}_{\alpha,1,1}(p_0 : q) &= -\log \frac{\int_{\mathcal{X}} p_0(x) q(x) \mathrm{d}x}{\left(\int_{\mathcal{X}} p_0(x)^{\alpha} \mathrm{d}x\right)^{1/\alpha} \left(\int_{\mathcal{X}} q(x)^{\beta} \mathrm{d}x\right)^{1/\beta}} \\ &= -\log \frac{\int_{\mathcal{X}} p(x)^{\frac{1}{\alpha-1}} q(x) \mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\frac{\alpha}{\alpha-1}} \mathrm{d}x\right)^{1/\alpha} \left(\int_{\mathcal{X}} q(x)^{\beta} \mathrm{d}x\right)^{1/\beta}} \\ &= -\log \frac{\int_{\mathcal{X}} p(x)^{\frac{\beta}{\alpha}} q(x) \mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\beta} \mathrm{d}x\right)^{1/\alpha} \left(\int_{\mathcal{X}} q(x)^{\beta} \mathrm{d}x\right)^{1/\beta}} = D^{\mathtt{H}}_{\alpha,\beta}(p : q). \end{aligned} \tag{87}$$

This means that the pre-aim technique of HPD is equivalent to HD $D^{\mathtt{H}}_{\alpha,\gamma}$ when we set $\gamma = \beta$.

As an alternative implementation of pre-aim, since $D^{\mathtt{H}}_{\alpha,1,1}(p : p^{\alpha-1}) = 0$, a proper divergence between $p(x)$ and $q(x)$ can be constructed by measuring:

$$D^{\mathtt{H}}_{\alpha,1,1}(q : p^{\alpha-1}) = -\log \frac{\int_{\mathcal{X}} p(x)^{\frac{\alpha}{\beta}} q(x) \mathrm{d}x}{\left(\int_{\mathcal{X}} q(x)^{\alpha} \mathrm{d}x\right)^{1/\alpha} \left(\int_{\mathcal{X}} p(x)^{\alpha} \mathrm{d}x\right)^{1/\beta}} = D^{\mathtt{H}}_{\beta,\alpha}(p : q), \tag{88}$$

turning out again to belong to the class of HD.

In practice, HD as a bi-parametric family may be less used than HPD with pre-aim because of the difficulty in choosing the parameter γ and because that HD has a slightly more complicated expression. The family of HD connecting CS divergence with skew Bhattacharyya divergence [8] is nevertheless of theoretical importance.

- In manifold learning [44–47], it is an essential topic to align two category distributions $p_0(x)$ and $q(x)$ corresponding respectively to the input and output [47], both for learning and for performance evaluation. In this case, the dimensionality of the statistical manifold that encompasses $p_0(x)$ and $q(x)$ is so high that to preserve monotonically $p_0(x)$ in the resulting $q(x)$ is already a difficult non-linear optimization and could be sufficient for the application, while preserving perfectly the input $p_0(x)$ is not so meaningful because of the input noise. It is then much easier to define pseudo-divergences using inequalities which do not necessarily need to be proper with potentially more choices. On the other hand, projective divergences including Hölder divergences introduced in this work are more meaningful in manifold learning than KL divergence (which is widely used) because they give scale invariance of the probability densities, meaning that one can define positive similarities then directly align these similarities, which is guaranteed to be equivalent to aligning the corresponding distributions. This could potentially give unified perspectives in between the two approaches of similarity-based manifold learning [46] and the probabilistic approach [44].

Hölder-type inequalities have been generalized to sets [48] instead of pairs of objects and to positive functional spaces, as well [49]. We also note that some divergences like Csiszár f-divergences enjoy themselves Hölder-type inequalities [50].

We expect that these two novel parametric Hölder classes of statistical divergences and pseudo-divergences open up new insights and applications in statistics and information sciences. Furthermore, the framework to build divergences or pseudo-divergences from proper or improper biparametric inequalities [1] offers novel classes of divergences to study.

Reproducible source code is available online [51].

Acknowledgments: The authors gratefully thank the referees for their comments. Ke Sun is funded by King Abdullah University of Science and Technology (KAUST).

Author Contributions: Frank Nielsen discussed the seminal ideas with Ke Sun and Stéphane Marchand-Maillet. Frank Nielsen and Ke Sun contributed to the theoretical results as well as to the writing of the article. Ke Sun implemented the methods and performed the numerical experiments. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

D_α^{HS}	Hölder proper non-projective Scored-induced divergence [18]
$D_{\alpha,\sigma,\tau}^{\mathrm{H}}$	Hölder improper projective pseudo-divergence (new)
$D_{\alpha,\gamma}^{\mathrm{H}}$	Hölder proper projective divergence (new)
D_α^{HE}	Hölder proper projective escort divergence (new)
KL	Kullback-Leibler divergence [10]
CS	Cauchy–Schwarz divergence [2]
B	Bhattacharyya distance [25]
$B_{\frac{1}{\alpha}}$	skew Bhattacharyya distance [8]
D_γ	γ-divergence (score-induced) [11]
p_α^E, q_β^E	escort distributions
α, β	Hölder conjugate pair of exponents: $\frac{1}{\alpha}+\frac{1}{\beta}=1$
$\bar{\alpha}, \beta$	Hölder conjugate exponent: $\bar{\alpha}=\beta=\frac{\alpha}{\alpha-1}$
θ_p, θ_q	natural parameters of exponential family distributions
$\mathcal{X}$	support of distributions
μ	Lebesgue measure
$L^\gamma(\mathcal{X},\mu)$	Lebesgue space of functions f such that $\int_{\mathcal{X}} \lvert f(x)\rvert^\gamma \mathrm{d}x < \infty$

Appendix A. Proof of Hölder Ordinary and Reverse Inequalities

We extend the proof ([52], p. 78) to prove both the (ordinary or forward) Hölder inequality and the reverse Hölder inequality.

Proof. First, let us observe that $-\log(x)$ is strictly convex on $(0,+\infty)$ since $(-\log(x))''=\frac{1}{x^2}$. It follows that for $0<a<1$ that:

$$-\log(ax_1+(1-a)x_2) \leq -a\log(x_1)-(1-a)\log(x_2), \tag{A1}$$

where the equality holds iff $x_1=x_2$.

Conversely, when $a<0$ or $a>1$, we have:

$$-\log(ax_1+(1-a)x_2) \geq -a\log(x_1)-(1-a)\log(x_2), \tag{A2}$$

where the equality holds iff $x_1=x_2$.

Equivalently, we can write these two inequalities as follows:

$$\begin{cases} x_1^a x_2^{1-a} & \leq ax_1+(1-a)x_2 \quad (\text{if } 0<a<1); \\ x_1^a x_2^{1-a} & \geq ax_1+(1-a)x_2 \quad (\text{if } a<0 \text{ or } a>1), \end{cases} \tag{A3}$$

both of them are tight iff $x_1=x_2$.

Let P and Q be positive measures with Radon–Nikodym densities $p(x)>0$ and $q(x)>0$ be positive densities with respect to the reference Lebesgue measure μ. The densities are strictly greater than zero on the support $\mathcal{X}$. Plugging:

$$a=\frac{1}{\alpha}, \quad 1-a=\frac{1}{\beta}, \quad x_1=\frac{p(x)^\alpha}{\int_{\mathcal{X}} p(x)^\alpha \mathrm{d}x}, \quad x_2=\frac{q(x)^\beta}{\int_{\mathcal{X}} q(x)^\beta \mathrm{d}x}, \tag{A4}$$

into Equation (A3), we get:

$$\begin{cases} \frac{p(x)}{\left(\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x\right)^{1/\alpha}} \frac{q(x)}{\left(\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x\right)^{1/\beta}} \leq \frac{1}{\alpha}\frac{p(x)^{\alpha}}{\int p(x)^{\alpha}\mathrm{d}x} + \frac{1}{\beta}\frac{q(x)^{\beta}}{\int q(x)^{\beta}\mathrm{d}x} & \text{if } \alpha > 0 \text{ and } \beta > 0, \\ \frac{p(x)}{\left(\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x\right)^{1/\alpha}} \frac{q(x)}{\left(\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x\right)^{1/\beta}} \geq \frac{1}{\alpha}\frac{p(x)^{\alpha}}{\int p(x)^{\alpha}\mathrm{d}x} + \frac{1}{\beta}\frac{q(x)^{\beta}}{\int q(x)^{\beta}\mathrm{d}x} & \text{if } \alpha < 0 \text{ or } \beta < 0. \end{cases} \tag{A5}$$

Assume that $p(x)$ in $L_{\alpha}(\mathcal{X}, \mu)$ and $q(x)$ in $L_{\beta}(\mathcal{X}, \mu)$, so that both $\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x$ and $\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x$ converge. Integrate both sides on $\mathcal{X}$ to get:

$$\begin{cases} \frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x\right)^{1/\alpha}\left(\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x\right)^{1/\beta}} \leq 1 & \text{if } \alpha > 0 \text{ and } \beta > 0, \\ \frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x\right)^{1/\alpha}\left(\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x\right)^{1/\beta}} \geq 1 & \text{if } \alpha < 0 \text{ or } \beta < 0. \end{cases} \tag{A6}$$

The necessary and sufficient condition for equality is that:

$$\frac{p(x)^{\alpha}}{\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x} = \frac{q(x)^{\beta}}{\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x}, \tag{A7}$$

almost everywhere. That is, there exists a positive constant $\lambda > 0$, such that:

$$p(x)^{\alpha} = \lambda q(x)^{\beta}, \quad \lambda > 0, \quad \text{almost everywhere.} \tag{A8}$$

□

The Hölder conjugate exponents α and β satisfies $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. That is, $\beta = \frac{\alpha}{\alpha - 1}$. Thus, when $\alpha < 0$, we necessarily have $\beta > 0$, and vice versa.

We can unify these two straight and reverse Hölder inequalities into a single inequality by considering the sign of $\alpha\beta = \frac{\alpha^2}{\alpha - 1}$: We get the general Hölder inequality:

$$\mathrm{sign}(\alpha\beta)\frac{\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x}{\left(\int_{\mathcal{X}} p(x)^{\alpha}\mathrm{d}x\right)^{1/\alpha}\left(\int_{\mathcal{X}} q(x)^{\beta}\mathrm{d}x\right)^{1/\beta}} \geq \mathrm{sign}(\alpha\beta). \tag{A9}$$

When $\alpha = \beta = 2$, the Hölder inequality becomes the Cauchy–Schwarz inequality:

$$\int_{\mathcal{X}} p(x)q(x)\mathrm{d}x \leq \sqrt{\left(\int_{\mathcal{X}} p(x)^2\mathrm{d}x\right)\left(\int_{\mathcal{X}} q(x)^2\mathrm{d}x\right)}. \tag{A10}$$

Historically, Cauchy stated the discrete sum inequality in 1821, while Schwarz reported the integral form of the inequality in 1888.

References

1. Mitrinovic, D.S.; Pecaric, J.; Fink, A.M. *Classical and New Inequalities in Analysis*; Springer Science & Business Media: New York, NY, USA, 2013; Volume 61.
2. Budka, M.; Gabrys, B.; Musial, K. On accuracy of PDF divergence estimators and their applicability to representative data sampling. *Entropy* **2011**, *13*, 1229–1266.
3. Amari, S.I. *Information Geometry and Its Applications*; Applied Mathematical Sciences series; Springer: Tokyo, Japan, 2016.
4. Rao, C.R. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
5. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.

6. Nielsen, F.; Nock, R. Total Jensen divergences: Definition, properties and clustering. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 2016–2020.
7. Burbea, J.; Rao, C. On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Inf. Theory* **1982**, *28*, 489–495.
8. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466.
9. Gneiting, T.; Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
10. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
11. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.
12. Nielsen, F.; Nock, R. Patch Matching with Polynomial Exponential Families and Projective Divergences. In Proceedings of the 9th International Conference on Similarity Search and Applications, Tokyo, Japan, 24–26 October 2016; pp. 109–116.
13. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.
14. Zhang, J. Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy* **2013**, *15*, 5384–5418.
15. Nielsen, F.; Nock, R. A closed-form expression for the Sharma–Mittal entropy of exponential families. *J. Phys. A Math. Theor.* **2011**, *45*, 032003.
16. De Souza, D.C.; Vigelis, R.F.; Cavalcante, C.C. Geometry Induced by a Generalization of Rényi Divergence. *Entropy* **2016**, *18*, 407.
17. Kanamori, T.; Fujisawa, H. Affine invariant divergences associated with proper composite scoring rules and their applications. *Bernoulli* **2014**, *20*, 2278–2304.
18. Kanamori, T. Scale-invariant divergences for density functions. *Entropy* **2014**, *16*, 2611–2628.
19. Nielsen, F.; Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv* **2009**, arXiv:0911.4863.
20. Rogers, L.J. An extension of a certain theorem in inequalities. *Messenger Math.* **1888**, *17*, 145–150.
21. Holder, O.L. Über einen Mittelwertssatz. *Nachr. Akad. Wiss. Gottingen Math. Phys. Kl.* **1889**, *44*, 38–47.
22. Hasanbelliu, E.; Giraldo, L.S.; Principe, J.C. Information theoretic shape matching. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2436–2451.
23. Nielsen, F. Closed-form information-theoretic divergences for statistical mixtures. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, 11–15 November 2012; pp. 1723–1726.
24. Zhang, J. Reference duality and representation duality in information geometry. *Am. Inst. Phys. Conf. Ser.* **2015**, *1641*, 130–146.
25. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *35*, 99–109.
26. Srivastava, A.; Jermyn, I.; Joshi, S. Riemannian analysis of probability density functions with applications in vision. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
27. Nielsen, F.; Nock, R. On the Chi Square and Higher-Order Chi Distances for Approximating f-Divergences. *IEEE Signal Process. Lett.* **2014**, *1*, 10–13.
28. Nielsen, F.; Nock, R. Skew Jensen-Bregman Voronoi diagrams. In *Transactions on Computational Science XIV*; Springer: New York, NY, USA, 2011; pp. 102–128.
29. Nielsen, F.; Sun, K. Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities. *Entropy* **2016**, *18*, 442.
30. Notsu, A.; Komori, O.; Eguchi, S. Spontaneous clustering via minimum gamma-divergence. *Neural Comput.* **2014**, *26*, 421–448.
31. Rigazio, L.; Tsakam, B.; Junqua, J.C. An optimal Bhattacharyya centroid algorithm for Gaussian clustering with applications in automatic speech recognition. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 5–9 June 2000; Volume 3, pp. 1599–1602.
32. Davis, J.V.; Dhillon, I.S. Differential Entropic Clustering of Multivariate Gaussians. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2006; pp. 337–344.

33. Nielsen, F.; Nock, R. Clustering multivariate normal distributions. In *Emerging Trends in Visual Computing*; Springer: New York, NY, USA, 2009; pp. 164–174.
34. Allamigeon, X.; Gaubert, S.; Goubault, E.; Putot, S.; Stott, N. A scalable algebraic method to infer quadratic invariants of switched systems. In Proceedings of the 12th International Conference on Embedded Software, Amsterdam, The Netherlands, 4–9 October 2015; pp. 75–84.
35. Sun, D.L.; Févotte, C. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 6201–6205.
36. Banerjee, A.; Dhillon, I.S.; Ghosh, J.; Sra, S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.* **2005**, *6*, 1345–1382.
37. Gopal, S.; Yang, Y. Von Mises-Fisher Clustering Models. *J. Mach. Learn. Res.* **2014**, *32*, 154–162.
38. Rami, H.; Belmerhnia, L.; Drissi El Maliani, A.; El Hassouni, M. Texture Retrieval Using Mixtures of Generalized Gaussian Distribution and Cauchy-Schwarz Divergence in Wavelet Domain. *Image Commun.* **2016**, *42*, 45–58.
39. Bunte, K.; Haase, S.; Biehl, M.; Villmann, T. Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing* **2012**, *90*, 23–45.
40. Villmann, T.; Haase, S. Divergence-based vector quantization. *Neural Comput.* **2011**, *23*, 1343–1392.
41. Huang, J.B.; Ahuja, N. Saliency detection via divergence analysis: A unified perspective. In Proceedings of the 2012 21st International Conference on Pattern Recognition (ICPR), Tsukuba, Japan, 11–15 November 2012; pp. 2748–2751.
42. Pardo, L. *Statistical Inference Based on Divergence Measures*; CRC Press: Abingdon, UK, 2005.
43. Basu, A.; Shioya, H.; Park, C. *Statistical Inference: The Minimum Distance Approach*; CRC Press: Abingdon, UK, 2011.
44. Hinton, G.E.; Roweis, S.T. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems 15 (NIPS)*; MIT Press: Vancouver, BC, Canada, 2002; pp. 833–840.
45. Maaten, L.V.D.; Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
46. Carreira-Perpiñán, M.Á. The Elastic Embedding Algorithm for Dimensionality Reduction. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 21–25 June 2010; pp. 167–174.
47. Sun, K.; Marchand-Maillet, S. An Information Geometry of Statistical Manifold Learning. In Proceedings of the International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 1–9.
48. Cheung, W.S. Generalizations of Hölder's inequality. *Int. J. Math. Math. Sci.* **2001**, *26*, 7–10.
49. Hazewinkel, M. *Encyclopedia of Mathematics*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001.
50. Chen, G.S.; Shi, X.J. Generalizations of Hölder inequalities for Csiszár's f-divergence. *J. Inequal. Appl.* **2013**, *2013*, 151.
51. Nielsen, F.; Sun, K.; Marchand-Maillet, S. On Hölder Projective Divergences. 2017. Available online: https://www.lix.polytechnique.fr/~nielsen/HPD/ (accessed on 16 March 2017).
52. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.