On information projections between multivariate elliptical and location-scale families

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

Abstract

We study information projections with respect to statistical $f$-divergences between any two location-scale families. We consider a multivariate generalization of the location-scale families which includes the elliptical and the spherical subfamilies. By using the action of the multivariate location-scale group, we show how to reduce the calculation of $f$-divergences between any two location-scale densities to canonical settings involving standard densities, and derive thereof fast Monte Carlo estimators of $f$-divergences with good properties. Finally, we prove that the minimum $f$-divergence between a prescribed density of a location-scale family and another location-scale family is independent of the prescribed location-scale parameter. We interpret geometrically this property.

Keywords: Information geometry, information projection, $f$-divergence, Kullback-Leibler divergence, location-scale family, and location-scale group.

1 Introduction

The concept of an information projection was first studied in information theory by Csiszár [9, 11] as the minimization of the Kullback-Leibler divergence (also called $I$-divergence) between a prescribed measure and a set of measures: When the minimum is unique, it is called the $I$-projection [10]. In information geometry [11, 25], the geometric study of information projections (e.g., conditions for uniqueness) is investigated as the geodesic projection with respect to an affine connection of a probability measure point onto a statistical submanifold [24] with orthogonality defined with respect to the Fisher-Rao metric. In this work, we consider information projections with respect to statistical $f$-divergences [8] when both the prescribed distribution and the subspace to project the distribution onto are multivariate generalizations of location-scale families which include the elliptical families and the spherical subfamilies.

We outline the paper with its main contributions as follows:

We first describe the multivariate generalization of location-scale families and introduce the multivariate location-scale group in §2. We then report several results for calculating the $f$-divergences between two densities of potentially different location-scale families in §3. Invariance of the $f$-divergences with respect to the action of the location-scale group (Theorem 1), calculations of the $f$-divergences by reduction to canonical settings (Corollary 1 exemplified for the Kullback-Leibler divergence in Corollary 3 and instantiated for the multivariate normal distributions), and invariance of $f$-divergences to scale for scale families (Corollary 2). In §4 we build efficient Monte Carlo
estimators with good properties to estimate the \( f \)-divergences between location-scale families when it is not calculable in closed-form. Finally, equipped with these preliminary results, we study in §5 the information projections of a prescribed distribution belonging to one location-scale family onto another location-scale family (Theorem 2), and we interpret geometrically these results.

## 2 Location-scale families and the location-scale group

### 2.1 Univariate location-scale families

Let \( X \sim p \) be a continuous random variable with cumulative distribution function (CDF) \( F_X \) and probability density function (PDF) \( p_X(x) \) defined on the support \( \mathcal{X} \). A location-scale random variable \( Y \overset{d}{=} l + sX \) (equality in distribution) for location parameter \( l \) and scale parameter \( s > 0 \) has CDF \( F_Y(y) = F_X \left( \frac{y - l}{s} \right) \) and PDF \( p_Y(y) = \frac{1}{s} p_X \left( \frac{y - l}{s} \right) \). Let \( p_{l,s}(x) := \frac{1}{s} p_X \left( \frac{x - l}{s} \right) \) denote the location-scale density for parameter \((l, s)\). The density \( p = p_{0,1} \) is called the standard density of the location-scale family. The location-scale parameter space of the location-scale family \( F_p = \{ p_{l,s}(x) : l \in \mathbb{R}, s > 0 \} \) is the upper plane \( \mathbb{H} = \mathbb{R} \times \mathbb{R}^{++} \).

**Example 1.** For example, the family of univariate normal distributions:

\[
\mathcal{N} := \left\{ p_{\mu,\sigma}^N(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^{++} \right\}
\]

is a location-scale family for the standard density \( p^N(x) := \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x^2) \) defined on \( \mathcal{X} = \mathbb{R} \) with location parameter \( l = \mu \) (the normal mean) and scale parameter \( s = \sigma > 0 \) (the normal standard deviation).

**Example 2.** Another example is the location-scale family of univariate Cauchy distributions:

\[
\mathcal{C} := \left\{ p_{l,s}^C(x) = \frac{1}{\pi s \left( 1 + \left( \frac{x - l}{s} \right)^2 \right)} : (l, s) \in \mathbb{R} \times \mathbb{R}^{++} \right\},
\]

with standard density \( p^C(x) := \frac{1}{\pi (1 + x^2)} \).

When \( E[p] \) is finite, we have \( E[Y] = l + sE[X] \), and when \( E[p^2] \) is finite, we have \( \sigma[Y] = \sqrt{E[(Y - E[Y])^2]} = s\sigma[X] \). Thus if we assume that the standard density \( p \) is such that \( E_p[X] = 0 \) and \( E_p[X^2] = 1 \) (i.e., \( p \) has unit variance), then the random variable \( Y \overset{d}{=} \mu + sX \) has mean \( E[Y] = \mu \) and standard deviation \( \sigma(Y) = \sqrt{E[(Y - \mu)^2]} = \sigma \). In the remainder, we do not use the \((\mu, \sigma)\) parameterization of location-scale families but the \((l, s)\) parameterization in order to be more general and consistent with the description of the multivariate location-scale families.

A location family is a family of densities \( \mathcal{L}_p = \{ p_l(x) = p(x - l) : l \in \mathbb{R} \} \). For example, the location family of shifted unit distributions with standard density \( p(x) = 1 \) on \( \mathcal{X} = [0, 1] \) is a location family. A location family can be obtained as a subfamily of a location-scale family \( \mathcal{F}_p \) by prescribing a scale \( s_0 > 0 \). For example, the family of normal distributions with unit variance is a location family, a subfamily of the normal location-scale family.

A scale family is a family of densities \( \mathcal{S}_p = \{ p_s(x) = \frac{1}{s} p \left( \frac{x}{s} \right) : s \in \mathbb{R}^{++} \} \). For example, the family of Rayleigh distributions \( \mathcal{R} := \{ \frac{x}{\sigma^2} \exp(-\frac{x^2}{2\sigma^2}) \} \) defined on the support \( \mathcal{X} = \mathbb{R}^{++} \) is a scale
family with standard density \( p^R(x) := x \exp(-\frac{x^2}{2}) \) and scale parameter \( s = \sigma^2 \). A scale family can be obtained as a subfamily of a location-scale family by prescribing a location \( l_0 \in \mathcal{X} \).

A location-scale family is said regular when its Fisher information matrix is positive-definite and finite. The location family induced by the uniform standard density on \([0, 1]\) is not a regular family since its Fisher information is infinite [14]. In the remainder, we consider regular location-scale families.

The Fisher-Rao geometry of location-scale families and its Riemannian distance [15, 16, 18] is recalled in Appendix A. The \( \alpha \)-geometry [11] of location-scale families have been studied in [22] who investigated the \( \alpha \)-geometry of univariate elliptical distributions with densities: \( \frac{1}{s} h \left( \frac{x - l}{s} \right)^2 \) for \((l, s) \in \mathbb{R} \times \mathbb{R}^+\). Thus by defining \( p(x) = h(x^2) \), we can convert any univariate elliptical distribution to a corresponding location-scale distribution (but a location-scale family is not necessarily an elliptical family because \( h(u) = p(\sqrt{u}) \) may not be properly defined for \( u < 0 \)). In particular, the \( \alpha \)-geometry of the Cauchy family is shown to be independent of \( \alpha \) (and never yielding a dually flat space [22]): Its conformal flattening into a dually flat geometry with applications to the construction of Voronoi diagrams has been studied in [26].

The location-scale parameter space \( \mathbb{H} \) form a group \( G = (\mathbb{H}, \cdot, \text{id}) \), called the location-scale group. An element \( g_{l,s} \in G \) acts \((\circ)\) on the standard density \( p(x) \) as follows:

\[
g_{l,s} \circ p(x) := \frac{1}{s} p \left( \frac{x - l}{s} \right). \tag{3}\]

The identity element is \( \text{id} = g_{0,1} \) since \( g_{0,1} \circ p = p \), and the group binary associative operation ‘ \( \cdot \)’ is retrieved from the group action as follows:

\[
g_{l_2,s_2} \cdot g_{l_1,s_1} \circ p = g_{l_2,s_2} \circ \left( \frac{1}{s_1} p \left( \frac{x - l_1}{s_1} \right) \right), \tag{4}\]

\[
= \frac{1}{s_1s_2} p \left( \frac{x - l_2 - l_1}{s_2} \right), \tag{5}\]

\[
=: g_{l_1,s_1}^{-1} \circ p, \tag{6}\]

with \( g_{l_1,s_1} \in G \) and \( l_{12} = s_2l_1 + l_2 \) and \( s_{12} = s_1s_2 \). The group inverse element is \( g_{l,s}^{-1} = g_{-\frac{l}{s}, \frac{1}{s}} \) which is obtained by solving \( g_{l,s} \cdot g_{l',s'} = \text{id} \): We \( l + s' = 0 \) and \( ss' = 1 \) solves as \( l' = -\frac{l}{s} \) and \( s' = \frac{1}{s} \). The orbit of the action of the location-scale group on the standard density \( p \) defines the location-scale family \( \mathcal{F}_p \):

\[
\mathcal{F}_p = G \circ p := \{g \circ p : \forall g \in G\}. \tag{7}
\]

The elements of the location-scale group can be represented using \( 2 \times 2 \) matrices (representation theory): Each group element \( g := g_{l,s} \) is represented by a corresponding matrix \( M_{l,s} := \begin{bmatrix} s & l \\ 0 & 1 \end{bmatrix} \). This matrix representation of elements yields the location-scale matrix group \((G, \times, I)\) with:

\[
G = \left\{ M_{l,s} = \begin{bmatrix} s & l \\ 0 & 1 \end{bmatrix} : (l, s) \in \mathbb{R} \times \mathbb{R}^+ \right\}, \tag{8}
\]

where the matrix group operation \( \times \) is the matrix multiplication, the identity element the \( 2 \times 2 \) matrix identity \( M_{0,0} = M_{0,1} = I \), and the inverse operation the matrix inverse:

\[
M_{g^{-1}} = (M_g)^{-1} = \begin{bmatrix} \frac{1}{s} & -\frac{l}{s} \\ 0 & 1 \end{bmatrix}. \tag{9}
\]

3
The location-scale group is a Lie matrix group \( \mathbb{R} \) (i.e., a “continuous group” modeled as a manifold) which acts transitively on the sample space. The location-scale group is non-abelian (i.e., non-commutative) because \( g_1 \cdot g_2 = g_1 + t_{s_1,s_2} \neq g_2 \cdot g_1 \) (since \( g_2 \cdot g_1 = g_2 + t_{s_1,s_2} \)). However the location subgroups and the scale subgroups are abelian groups. Representing elements by matrices is handy to prove basic properties: For example, we can prove easily that \((g_1 \cdot g_2)^{-1} = g_2^{-1} \cdot g_1^{-1}\) since Jacobians of \( g_1 \) and \( g_2 \) do not commute.

\[ (M_{g_1} \times M_{g_2})^{-1} = M_{g_2}^{-1} \times M_{g_1}^{-1} = M_{g_2^{-1} \cdot g_1^{-1}}. \]  

### 2.2 Multivariate location-scale families: Location-positive families

Let \( \mathcal{P}(X) \) denote the set of probability density functions with support \( X \).

We begin by first recalling the relationships between the PDFs of two continuous \( d \)-dimensional random variables \( X = (X_1, \ldots, X_d) \sim p_X \) and \( Y = t(X) = (t_1(X), \ldots, t_d(X)) \sim p_Y \) for a differentiable and invertible function \( t \) with non-singular Jacobian (i.e., \( |\text{Jac}_t(x)| \neq 0, \forall x \in X \)).

The location-scale group is a Lie matrix group \( \mathbb{R} \) (i.e., a “continuous group” modeled as a manifold) which acts transitively on the sample space. The location-scale group is non-abelian (i.e., non-commutative) because \( g_1 \cdot g_2 = g_1 + t_{s_1,s_2} \neq g_2 \cdot g_1 \) (since \( g_2 \cdot g_1 = g_2 + t_{s_1,s_2} \)). However the location subgroups and the scale subgroups are abelian groups. Representing elements by matrices is handy to prove basic properties: For example, we can prove easily that \((g_1 \cdot g_2)^{-1} = g_2^{-1} \cdot g_1^{-1}\) since Jacobians of \( g_1 \) and \( g_2 \) do not commute.

\[ (M_{g_1} \times M_{g_2})^{-1} = M_{g_2}^{-1} \times M_{g_1}^{-1} = M_{g_2^{-1} \cdot g_1^{-1}}. \]  

Furthermore, we have the following identity:

\[ |\text{Jac}_t(x)| \times |\text{Jac}_t^{-1}(y)| = |\text{Jac}_t(x) \times \text{Jac}_t^{-1}(y)| = |I| = 1, \]  

where \( I \) denotes the \( d \times d \) identity matrix.

For sanity checks, we verify that we have:

\[ p_X(x) = |\text{Jac}_t(x)| \times p_Y(t(x)) = |\text{Jac}_t(x)| \times p_Y(y), \]  
\[ p_Y(y) = |\text{Jac}_t^{-1}(y)| \times p_X(t^{-1}(y)) = |\text{Jac}_t^{-1}(y)| \times p_X(x). \]  

Furthermore, we have the following identity:

\[ |\text{Jac}_t(x)| \times |\text{Jac}_t^{-1}(y)| = |\text{Jac}_t(x) \times \text{Jac}_t^{-1}(y)| = |I| = 1, \]  

where \( I \) denotes the \( d \times d \) identity matrix.

For sanity checks, we verify that we have:

\[ p_X(x) = |\text{Jac}_t(x)| \times p_Y(t(x)) = |\text{Jac}_t(x)| \times |\text{Jac}_t^{-1}(y)| \times p_X(t^{-1}(y)), \]  
\[ = |\text{Jac}_t(x) \times \text{Jac}_t^{-1}(y)| \times p_X(x) = |I|p_X(x) = p_X(x), \]  

since \( \text{Jac}_t(x) \times \text{Jac}_t^{-1}(y) = I \).

Let \( X \) be a \( d \)-dimensional multivariate random variable, and let \( Y \overset{d}{=} PX + l \) for \( P > 0 \) a positive-definite \( d \times d \) matrix playing the role of the “multidimensional scale” parameter, and \( l \in \mathbb{R}^d \) a location parameter. Then using Eq. 13 with \( Y \overset{d}{=} t_{l,P}(X) = PX + l \) and \( X \overset{d}{=} t^{-1}_{l,P}(Y) = P^{-1}(Y - l) \), we find the density of \( p_{l,P} \) of continuous random distribution \( Y \) as follows:

\[ p_{l,P}(y) = |\text{Jac}_{t^{-1}_{l,P}}(y)| \times p_X(t^{-1}_{l,P}(y)) = |\text{Jac}_{t^{-1}_{l,P}}(y)| \times p_X(x), \]  
\[ = |P^{-1}| p_X(y - l), \]  

where \( p_X \) denotes the standard density since \( \text{Jac}_t^{-1}(y) = P^{-1} \). The space of multivariate location-scale parameters \( (l,P) \) is \( \mathbb{R}_+^d = \mathbb{R}^d \times \mathbb{P}_{++} \), where \( \mathbb{P}_{++} \) denotes the open cone of positive-definite matrices. Observe that by embedding \((l,P)\) as \((\text{diag}(l_1, \ldots, l_d), P)\) (where
\( M = \text{diag}(l_1, \ldots, l_d) \) denotes the diagonal matrix with \( M_{ii} = l_i \), we obtain a parameter domain which is a subspace of the Siegel upper plane \([27] \text{Sym}(\mathbb{R}, d) \times \mathbb{P}_{++}\), where \( \text{Sym}(\mathbb{R}, d) \) denotes the space of symmetric \( d \times d \) matrices.

When \( d = 1 \) and \( P = s \), we have \( Y \equiv t_{l,s}(X) = sX + l \), \( X \equiv t_{l,s}^{-1}(Y) = \frac{1}{s}(Y - l) \) and we recover the univariate location-scale densities \( p_{l,s}(y) = \frac{1}{s}p \left( \frac{y - l}{s} \right) \).

We can define equivalently the density of a location-scale family by \( p_{l,P}(x) = |P|^{-1}p \left( P^{-1}(x - l) \right) \) since \( |P^{-1}| = |P|^{-1} \). Since \( P \) is a positive-definite matrix generalizing the position scalar in the location-scale group, we also call this multivariate generalization of the location-scale group, the location-positive group. Thus the location-positive families can be obtained as the action of the location-positive group on a prescribed density \( p \in \mathcal{P}(\mathbb{R}^d) \) (or \( \mathcal{P}(\mathbb{R}^d_{++}) \) for scale only families).

**Definition 1** (Multivariate location-scale/location-positive family). Let \( p \in \mathcal{P}(\mathbb{R}^d) \) be a probability density function on \( \mathbb{R}^d \). Then the multivariate location-scale family is:

\[
\mathcal{F}_p = \left\{ p_{l,P}(x) = |P|^{-1}p \left( P^{-1}(x - l) \right) : (l, P) \in \mathbb{R}^d \times \mathbb{P}_{++} \right\}.
\] (19)

For example, the family of multivariate normal distributions (MVNs) is a multivariate location-scale family where the standard PDF is:

\[
p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp \left( -\frac{1}{2} x^\top x \right).
\] (20)

Indeed, the covariance matrix \( \Sigma \) is a positive-definite matrix which admits a unique symmetric positive-definite square root matrix \( \Sigma^{\frac{1}{2}} \) (such that \( \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} = \Sigma \)). This symmetric square root matrix can be calculated from the eigendecomposition of \( \Sigma \) in cubic time \( O(d^3) \) as follows: Let \( \Sigma = V^\top \text{diag}(\lambda_1, \ldots, \lambda_d)V^{-1} \) denote the eigendecomposition where the \( \lambda_i \)'s are the positive real eigenvalues and \( V \) the matrix of column eigenvectors. Then \( \sqrt{\Sigma} = \Sigma^{\frac{1}{2}} = V \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})V^{-1} \), and \( \Sigma^{\frac{1}{2}} \Sigma^{\frac{1}{2}} = V \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})V^{-1}V \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d})V^{-1} = V^\top \text{diag}(\lambda_1, \ldots, \lambda_d)V^{-1} = \Sigma \) since \( V^{-1}V = I \). Notice that \( \sqrt{\Sigma} \) is a positive-definite matrix. We have:

\[
p_{\mu, \Sigma^{\frac{1}{2}}}(y) = \left| \Sigma^{-\frac{1}{2}} \right| p \left( \Sigma^{-\frac{1}{2}}(y - \mu) \right),
\] (21)

\[
= \frac{\left| \Sigma^{-\frac{1}{2}} \right|}{(2\pi)^{\frac{d}{2}}} \exp \left( -\frac{1}{2} (\Sigma^{-\frac{1}{2}}(y - \mu))^\top \Sigma^{-\frac{1}{2}}(y - \mu) \right),
\] (22)

\[
= \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{\left| \Sigma \right|}} \exp \left( -\frac{1}{2} (y - \mu)^\top \Sigma^{-1}(y - \mu) \right),
\] (23)

since \( \left| \Sigma^{-\frac{1}{2}} \right| = \frac{1}{\left| \Sigma^{\frac{1}{2}} \right|} = \frac{1}{\sqrt{\left| \Sigma \right|}} \) and \( (\Sigma^{-\frac{1}{2}}(y - \mu))^\top = (y - \mu)^\top \Sigma^{-\frac{1}{2}} \) since \( \Sigma = \Sigma^\top \) (and by using the matrix trace cyclic property). Eq. 23 recovers the multivariate normal density. It follows that if \( X \sim \mathcal{N}(\mu, \Sigma) \) then we have \( Y = \Sigma^{-\frac{1}{2}}(X - \mu) \sim \mathcal{N}(0, I) \).
Multivariate location-scale families include the elliptical families which have densities of the form [17]:

\[ p_{\mu,V}^{\text{ell}}(x) = |V|^{-\frac{1}{2}}h\left((x - \mu)^\top V^{-1}(x - \mu)\right), \]

where \( h \) is a profile function. Indeed, let \( P = V^{\frac{1}{2}} \) and \( \mu = l \) with \( p(x) = h(x^\top x) \). Then we have

\[
\begin{align*}
p_{\mu,V}^{\frac{1}{2}} &= |V|^{-\frac{1}{2}}h\left((V^{-\frac{1}{2}}(x - \mu))^\top (V^{-\frac{1}{2}}(x - \mu))\right), \\
&= |V|^{-\frac{1}{2}}h\left((x - \mu)^\top V^{-1}(x - \mu)\right) := p_{\mu,V}^{\text{ell}}(x).
\end{align*}
\]

Moreover, the elliptical families include the spherical subfamilies as a special case when \( P = I \), see [17]. Last, let us remark that some parametric families of distributions can be both interpreted as location-scale families and exponential families [2] (e.g., normal family, Rayleigh family, inverse Gaussian family, and gamma family).

The multivariate location-scale group \( G_d \) can be defined on the multivariate location-scale parameter space \( G_d = \mathbb{R}^d \times \mathbb{R}^d_+ \). The identity element is \( \text{id} = (0, I) \), the group operation is \( g_{l_2,P_2} g_{l_1,P_1} = g_{l_2 + P_2 l_1, P_2 P_1} \). This group operation rule can be found by the action of the location-scale group onto the standard density:

\[
g_{l_2,P_2} g_{l_1,P_1} \odot p(x) = |P_2|^{-1} |P_1|^{-1} p\left(P_1^{-1}(P_2^{-1}(x - l_2) - l_1)\right),
\]

\[
= (P_2 P_1)^{-1} p\left((P_2 P_1)^{-1} x - (P_2 P_1)^{-1} l_2 - P_1^{-1} l_1\right),
\]

\[
= (P_2 P_1)^{-1} p\left((P_2 P_1)^{-1} x - l_2 - P_2 l_1\right).
\]

The action of the multivariate location-scale group on a density \( p \) is given by:

\[
g_{l,P} \odot p := |P|^{-1} p\left(|P|^{-1}(x - l)\right).
\]

The multivariate location-scale family (i.e., set of location-scale models) is obtained by taking the group orbit of the standard density \( p \):

\[
\mathcal{F}_p = G_d \odot p.
\]

Thus the location-scale group \( (G_d, ., \text{id}) \) is represented by the location-scale matrix group \( (G_d, \times, I_{d+1}) \).

The corresponding multivariate location-scale block matrix group is the following set of \((d + 1) \times (d + 1)\) matrices:

\[
G_d = \left\{ M_{l,P} = \begin{bmatrix} P & l \\ 0_d^\top & 1 \end{bmatrix} : (l, P) \in \mathbb{R}^d \times \mathbb{R}^d_+ \right\},
\]

(32)

The inverse element \( g_{l,P}^{-1} = g_{-P^{-1}l,P^{-1}} \) can be found from the matrix inverse of \( M_{l,P} \). Indeed, we check that:

\[
\begin{bmatrix} P & l \\ 0_d^\top & 1 \end{bmatrix} \begin{bmatrix} P^{-1} & -P^{-1} l \\ 0_d^\top & 1 \end{bmatrix} = \begin{bmatrix} I & 0_d \\ 0_d^\top & 1 \end{bmatrix} = I_{d+1}.
\]

(33)

The matrix group multiplication is

\[
\begin{bmatrix} P_1 & l_1 \\ 0_d^\top & 1 \end{bmatrix} \times \begin{bmatrix} P_2 & l_2 \\ 0_d^\top & 1 \end{bmatrix} = \begin{bmatrix} P_1 P_2 & P_1 l_2 + l_1 \\ 0_d^\top & 1 \end{bmatrix}.
\]

(34)
The Fisher-Rao geometry and α-geometry of multivariate normal distributions was studied in [37] and is reviewed in [23]. More generally, Mitchell studied the α-geometry of elliptical families [21]. Ohara and Eguchi [32] studied some dually flat geometry of elliptical families. Warped Riemannian metrics have also been studied for location-scale families defined on a Riemannian manifold [30] (including the Euclidean manifold $\mathbb{R}^d$): For example, the family of $d$-dimensional isotropic normal distributions is a multivariate location family whose Fisher-Rao metric is a warped Riemannian metric.

3 Statistical divergences between location-scale densities

Let us consider the statistical $f$-divergences [8] $I_f$ between two continuous distributions $p$ and $q$ of $\mathbb{R}^d$:

$$I_f(p : q) = \int_{x \in \mathcal{X}} p(x) f \left( \frac{q(x)}{p(x)} \right) dx,$$

(35)

where $f$ is a convex function, strictly convex at 1, satisfying $f(1) = 0$. When the $f$-divergence generator is chosen to be $f(u) = -\log(u)$, we retrieve the Kullback-Leibler divergence (KLD):

$$D_{KL}(p : q) = \int p(x) \log \frac{p(x)}{q(x)} dx.$$

(36)

The reverse $f$-divergence $I_f^r(p : q) := I_f(q : p)$ is obtained for the conjugate generator $f^*(u) := uf(\frac{1}{u})$ (convex with $f^*(1) = 0$): $I_f^r(p : q) = I_f^r(p : q) = I_f(q : p)$.

Let $p = p_{0, I}$ and $q = q_{0, I}$ be the two standard PDFs with support $\mathbb{R}^d$ defining multivariate location-scale families $\mathcal{F}_p$ and $\mathcal{F}_q$, respectively. Let $p_{1, I}, p_2 \in \mathcal{F}_p$ and $q_{2, I} \in \mathcal{F}_q$.

We state the following group invariance theorem of the $f$-divergences:

**Theorem 1** (Invariance of $f$-divergences under the location-scale group). We have

$$I_f(g \odot p : g \odot q) = I_f(p : q)$$

for all $p, q \in \mathcal{P}(\mathbb{R}^d)$ and any $g = g_{l, p}$ in the multivariate location-scale group $G_d = \mathbb{R}^d \times \mathbb{P}^d_{++}$. 

_Proof._ We have

$$I_f(g \odot p : g \odot q) = \int |P|^{-1}p \left( |P|^{-1}(x - l) \right) \log \left( \frac{|P|^{-1}p \left( |P|^{-1}(x - l) \right)}{|P|^{-1}q \left( |P|^{-1}(x - l) \right)} \right) dx,$$

(37)

$$= \int p \left( |P|^{-1}(x - l) \right) \log \left( \frac{p(y)}{q(y)} \right) dy =: I_f(p : q),$$

(38)

after making a change of variable $y = |P|^{-1}(x - l)$ in the multiple integral $\int \ldots dx = \int \ldots \int dx_1 \ldots dx_d$ with $dy = |P|^{-1}dx$. This change of variable requires $\mathcal{X} = \mathbb{R}^d$ [20] and therefore $p, q \in \mathcal{P}(\mathbb{R}^d)$. Indeed, when the support of the PDFs are dependent of $(l, P)$ (e.g., a uniform distribution on a compact $\mathcal{K} \subset \mathbb{R}^d$), the KLD diverges and the Fisher information is infinite [14]. Thus we assume in the remainder that all location-scale families are regular. \qed
From Theorem 1 we get the following corollary:

**Corollary 1** (Canonical settings for $f$-divergences between location-scale distributions). The $f$-divergence between two regular location-scale densities is equivalent to the $f$-divergence between one standard location-scale density and another affinely shifted location-scale density:

$$I_f(p_{l_1,P_1} : q_{l_2,P_2}) = I_f\left(p : q_{P_1^{-1}(l_2-l_1),P_1^{-1}P_1} \right) = I_f\left(p_{P_2^{-1}(l_2-l_1),P_2^{-1}P_1} : q \right). \quad (39)$$

**Proof.** We give two proofs: A short indirect proof relying on Theorem 1 and a direct proof.

- Let $g_1 = g_{l_1,P_1}$ and $g_2 = g_{l_2,P_2}$ so that $p_{l_1,P_1} = p_{g_1}$ and $q_{l_2,P_2} = q_{g_2}$. Applying Theorem 1 with $g = g_1$, we have $I_f(p_{l_1,P_1} + g_2 \circ q) = I_f(g_1^{-1} \cdot g_1 \circ p : g_1^{-1} \cdot g_2 \circ q)$. Since $g_1^{-1} \cdot g_1 = \text{id}$ and $g_1^{-1} \cdot g_2 = g_{P_1^{-1}(l_2-l_1),P_1^{-1}P_2}$, we get $I_f(p_{l_1,P_1} : q_{l_2,P_2}) = I_f\left(p : q_{P_1^{-1}(l_2-l_1),P_1^{-1}P_2} \right)$. Similarly, applying Theorem 1 with $g = g_2$, we get $I_f(p_{l_1,P_1} : q_{l_2,P_2}) = I_f\left(p_{P_2^{-1}(l_2-l_1),P_2^{-1}P_1} : q \right)$ since $g_2^{-1} \cdot g_1 = g_{P_1^{-1}(l_2-l_1),P_2^{-1}P_1}$.

- The second direct proof makes the change of variable in $\star$ with $y = P_1^{-1}(x-l_1)$, $x = P_1y + l_1$, $dy = |P_1|^{-1}dx$ and $dx = |P_1|dy$, and uses the identity $|P_2|^{-1} = |P_1|^{-1}P_2|^{-1}$:

$$I_f(p_{l_1,P_1} : q_{l_2,P_2}) := \int_{\mathcal{X}} p_{l_1,P_1}(x) f\left(q_{l_2,P_2}(x) \right) dx, \quad (40)$$

$$= \int |P_1|^{-1} p \left(P_1^{-1}(x-l_1) \right) f\left| \frac{P_2^{-1}q\left(P_2^{-1}(x-l_2)\right)}{P_1^{-1}p\left(P_1^{-1}(x-l_1)\right)} \right| dx,$$

$$\star \int p(y) f\left| \frac{\left|P_2^{-1}q\left(P_2^{-1}(P_1y + \mu_1) - \mu_2\right)\right|}{|P_1|^{-1}p\left(P_1^{-1}(x-l_1)\right)} \right| dy,$$

$$= \int p(y) f\left| \frac{\left|P_1^{-1}P_2\right|^{-1}q\left(P_1^{-1}P_2\right)^{-1}(y-P_2^{-1}(l_2-l_1))}{p(y)} \right| dy, \quad (41)$$

$$= I_f\left(p : q_{P_2^{-1}(l_2-l_1),P_2^{-1}P_1} \right), \quad (42)$$

Using the conjugate generator $f\star(u)$, we get $I_f(p_{l_1,P_1} : q_{l_2,P_2}) = I_f\left(p_{P_2^{-1}(l_2-l_1),P_2^{-1}P_1} : q \right)$. 

Thus we obtain the scale invariance of the $f$-divergence between multivariate scale families (including zero-centered elliptical distributions):

**Corollary 2** (Scale invariance of $f$-divergences between scale densities). The $f$-divergence between multivariate scale densities $p_{P_1}$ and $q_{P_2}$ is scale-invariant: For all $\lambda > 0$: $I_f(p_{\lambda P_1} : p_{\lambda P_2}) = I_f(p_{P_1} : p_{P_2}) = I_f(p : q_{P_1^{-1}P_2}) = I_f(p_{P_2^{-1}P_1} : q)$. 

8
**Example 3.** Consider the Rayleigh scale family with \( X = \mathbb{R}_{++} \) and standard density \( p(x) = x \exp(-\frac{x^2}{2}) \). The KLD between two Rayleigh distributions is

\[
D_{KL}(p_{\sigma_1^2} : p_{\sigma_2^2}) = \frac{\sigma_1^2}{\sigma_2^2} - \log \left( \frac{\sigma_1^2}{\sigma_2^2} \right) - 1. \tag{43}
\]

We check that \( D_{KL}(g_\lambda \circ p_{\sigma_1^2} : g_\lambda \circ p_{\sigma_2^2}) = D_{KL}(p_{\sigma_1^2} : p_{\sigma_2^2}) \) since \( g_\lambda \circ p_{\sigma_2} = p_{\lambda \sigma_2} \) and \( \frac{\sigma_1^2}{\lambda \sigma_2^2} = \frac{\sigma_1^2}{\sigma_2^2} \). Similarly, the KLD between two univariate zero-centered normal distributions yields the same formula. In fact the Rayleigh distributions form an exponential family and the KLD amounts to a Bregman divergence which is the Itakura-Saito divergence \( D_{IS}(\theta_1 : \theta_2) := \frac{\sigma_1}{\sigma_2} - \log \frac{\sigma_1}{\sigma_2} - 1 \). We have \( D_{KL}(p_{\sigma_1^2} : p_{\sigma_2^2}) = D_{IS}(\theta_2 : \theta_1) \) with \( \theta_i = -\frac{1}{2 \sigma_i^2} \). See \([28]\) for details.

Let us instantiate the invariance property of Corollary 1 for the KLD. We get:

**Corollary 3 (KLD between location-scale densities).** The KLD between two regular location-scale densities is equivalent to the \( f \)-divergence between one standard location-scale density and another affinely shifted location-scale density:

\[
D_{KL}(p_{l_1,p_1} : q_{l_2,p_2}) = D_{KL} \left( p : q_{P_1^{-1}(l_2-l_1),P_1^{-1}P_2} \right) = D_{KL} \left( P_{P_2^{-1}(l_1-l_2),P_2^{-1}P_1} : q \right). \tag{44}
\]

Since KLD \( D_{KL}(p : q) \) amounts to the cross-entropy \( h^X(p : q) = -\int p(x) \log q(x) dx \) minus Shannon’s differential entropy \( h(p) = h^X(p : p) = -\int p(x) \log p(x) dx \), let us also report the formula for the cross-entropy/entropy under the action of a location-scale group element \( g = g_{l,P} \):

\[
h^X(g \circ p : g \circ q) = h^X(p : q) + \log |P|, \tag{45}
\]

\[
h(g \circ p) = h(p) + \log |P|. \tag{46}
\]

Thus \( D_{KL}(g \circ p : g \circ q) = h^X(g \circ p : g \circ q) - h(g \circ p) = h^X(p : q) - h(p) = D_{KL}(p : q) \).

Furthermore, we have:

\[
h^X(p_{l_1,p_1} : q_{l_2,p_2}) = h^X(p : q_{P_1^{-1}(l_2-l_1),P_1^{-1}P_2}) - \log |P_1|, \tag{47}
\]

\[
= h^X(p_{P_2^{-1}(l_1-l_2),P_2^{-1}P_1} : q) - \log |P_2|. \tag{48}
\]

Notice that it is well-known that the \( f \)-divergence between two continuous densities with full support in \( \mathbb{R}^d \) is independent of a diffeomorphism \([41]\) \( Y = \phi(X) \): That is, \( I_f(p_X(x) : q_X(x)) = I_f(p_Y(y) : q_Y(y)) \). The proof also makes use of a change of variable in a multiple integral and requires \([20]\) \( \mathcal{X} = \mathbb{R}^d \):

**Proposition 1 (Invariance of \( f \)-divergences).** Let \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^d \) be a diffeomorphism, \( p_X, q_X \in \mathcal{P}(\mathbb{R}^d) \) and \( Y = \phi(X) \). Then we have \( I_f(p_Y(y) : q_Y(y)) = I_f(p_X(x) : q_X(x)) \).
Proof. Let \( p_Y(y) = |\text{Jac}_{t^{-1}}(y)| \times p_X(t^{-1}(y)) \) and \( q_Y(y) = |\text{Jac}_{t^{-1}}(y)| \times q_X(t^{-1}(y)) \) with \( x = t^{-1}(y) \) and \( dx = |\text{Jac}_{t^{-1}}(y)| dy \). We have:

\[
I_f(p_Y : q_Y) = \int_{\mathbb{R}^d} p_Y(y) f \left( \frac{q_Y(y)}{p_Y(y)} \right) dx
\]

\[
= \int_{\mathbb{R}^d} |\text{Jac}_{t^{-1}}(y)| \times p_X(t^{-1}(y)) f \left( \frac{|\text{Jac}_{t^{-1}}(y)| \times q_X(t^{-1}(y))}{|\text{Jac}_{t^{-1}}(y)| \times p_X(t^{-1}(y))} \right) dy,
\]

\[
= \int_{\mathbb{R}^d} p_X(x) f \left( \frac{q_X(x)}{p_X(x)} \right) dx =: I_f(p_X : q_X).
\]

Letting \( Y = PX + l \), \( p_Y = g_{l,P} \circ p_X \) and \( q_Y = g_{l,P} \circ q_X \), we get \( I_f(g_{l,P} \circ p_X : g_{l,P} \circ q_X) = I_f(p_X : q_X) \).

Example 4. Consider the family of log-normal distributions such that if \( X \sim \mathcal{N}(\mu, \sigma) \) then \( Y = \exp(X) \) follows a log-normal distribution \( \mathcal{LN}(\mu, \sigma) \) with probability density function:

\[
p^{\mathcal{LN}}_{\mu,\sigma}(x) := \frac{1}{x \sigma \sqrt{2\pi}} \exp \left( -\frac{(\ln x - \mu)^2}{2\sigma^2} \right),
\]

for \( x \in X = (0, \infty) \). Reciprocally, if \( X \sim \mathcal{LN}(\mu, \sigma) \) then \( Y = \log(X) \) follows a normal distribution \( \mathcal{N}(\mu, \sigma) \). It follows from Proposition 3 that the \( f \)-divergence \( I_f(p^{\mathcal{LN}}_{\mu_1,\sigma_1} : p^{\mathcal{LN}}_{\mu_2,\sigma_2}) = I_f(p^\mathcal{N}_{\mu_1,\sigma_1} : p^\mathcal{N}_{\mu_2,\sigma_2}) \) (see also [13] for the matching formula of the Kullback-Leibler divergence).

We can define the \( f \)-\textit{mutual information} between two random variables \( X \) and \( Y \) as

\[
\text{MI}_f(X;Y) := I_f(p_{(X,Y)} : p_X p_Y).
\]

Whenever \( p_{(X,Y)} = p_X p_Y \), we say that random variable \( X \) is independent to random variable \( Y \), and the \( f \)-mutual information is zero: \( \text{MI}_f(X;Y) = 0 \). We have the following invariance of the mutual information:

Proposition 2 (Invariance of \( f \)-mutual information). For any invertible and differentiable transformations \( t_1 \) and \( t_2 \) from \( \mathbb{R}^d \) to \( \mathbb{R}^d \), we have \( \text{MI}_f(t_1(X_1);t_2(X_2)) = \text{MI}_f(X_1 : X_2) \).

Proof. Let \( Y_1 = t_1(X_1) \) and \( Y_2 = t_2(X_2) \). We have the joint density \( p_{(Y_1,Y_2)}(y_1,y_2) = |\text{Jac}_{t_1}^{-1}(y_1)| |\text{Jac}_{t_2}^{-1}(y_2)| p_{(X_1,X_2)}(x_1,x_2) \) and the marginals \( p_{Y_1}(y_1) = |\text{Jac}_{t_1}^{-1}(y_1)| p_{X_1}(x_1) \) and \( p_{Y_2}(y_2) = |\text{Jac}_{t_2}^{-1}(y_2)| p_{X_2}(x_2) \). It follows that \( \frac{p_{Y_1}(y_1)p_{Y_2}(y_2)}{p_{(Y_1,Y_2)}(y_1,y_2)} = \frac{p_{X_1}(x_1)p_{X_2}(x_2)}{p_{(X_1,X_2)}(x_1,x_2)} \). The \( f \)-mutual information \( \text{MI}_f(t_1(X_1);t_2(X_2)) \) rewrites as:

\[
\text{MI}_f(t_1(X_1);t_2(X_2)) = \int_{y_1} \int_{y_2} p_{(Y_1,Y_2)}(y_1,y_2) f \left( \frac{p_{Y_1}(y_1)p_{Y_2}(y_2)}{p_{(Y_1,Y_2)}(y_1,y_2)} \right) dy_1 dy_2,
\]

\[
= \int_{y_1} \int_{y_2} p_{(Y_1,Y_2)}(y_1,y_2) f \left( \frac{p_{X_1}(x_1)p_{X_2}(x_2)}{p_{(X_1,X_2)}(x_1,x_2)} \right) dy_1 dy_2.
\]
Using two changes of variables $x_1 = t_1^{-1}(y_1)$ and $x_2 = t_2^{-1}(x_2)$ with $|\text{Jac}_{t_1}(y_1)| \, dy_1 = dx_1$ and $|\text{Jac}_{t_2}(y_2)| \, dy_2 = dx_2$, we have:

$$p(y_1, y_2)(y_1, y_2)dy_1dy_2 = |\text{Jac}_{t_1}(y_1)| \, |\text{Jac}_{t_2}(y_2)| \, p(x_1, x_2)(x_1, x_2)dy_1dy_2, \quad (56)$$

$$\Rightarrow p(x_1, x_2)(x_1, x_2)dx_1dx_2. \quad (57)$$

Thus we have Eq. [55] which rewrites as:

$$\text{MI}_f(t_1(X_1); t_2(X_2)) = \int_{x_1} \int_{x_2} p(x_1, x_2)(x_1, x_2)f \left( \frac{p_x(x_1)p_x(x_2)}{p(x_1, x_2)(x_1, x_2)} \right) dx_1dx_2, \quad (58)$$

$$\Rightarrow \text{MI}_f(X_1 : X_2). \quad (59)$$

Notice that for the change of variables we require to have both the joint densities and the marginal densities to be defined on the full support $\mathbb{R}^d$.

Let us illustrate the formula of Eq. [44] in the following example:

**Example 5.** The KLD between the standard normal $p^N$ and a normal $p^N_{\mu, \sigma} = p_{\mu, \sigma}$ is

$$D_{KL}(p^N : p^N_{\mu, \sigma}) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2} \left( \frac{1}{\sigma^2} - \log \frac{1}{\sigma^2} - 1 \right) \quad (60).$$

From this formula, we recover the generic KLD formula between two normal distributions by plugging $\sigma = \frac{\sigma_2}{\sigma_1}$ and $\mu = \frac{\mu_2 - \mu_1}{\sigma_1}$:

$$D_{KL}(p^N_{\mu_1, \sigma_1} : p^N_{\mu_2, \sigma_2}) = D_{KL}(p^N : p^N_{\mu_2 - \mu_1, \frac{\sigma_2}{\sigma_1}}), \quad (61)$$

$$= \frac{(\mu_2 - \mu_1)^2}{2\sigma_2^2} + \frac{1}{2} \left( \frac{\sigma_1^2}{\sigma_2^2} - \log \frac{\sigma_1^2}{\sigma_2^2} - 1 \right). \quad (62)$$

Equivalently, we could also have used the canonical formula:

$$D_{KL}(p^N_{\mu_2, \sigma_2} : p^N) = \frac{1}{2} (\sigma^2 + \mu^2 - 1 - \log \sigma^2), \quad (63)$$

and then retrieve the ordinary formula as follows:

$$D_{KL}(p^N : p^N_{\mu, \sigma}) = D_{KL}(p^N_{\mu_2 - \mu_1, \frac{\sigma_1}{\sigma_2}} : p^N), \quad (64)$$

$$= \frac{(\mu_2 - \mu_1)^2}{2\sigma_2^2} + \frac{1}{2} \left( \frac{\sigma_1^2}{\sigma_2^2} - \log \frac{\sigma_1^2}{\sigma_2^2} - 1 \right). \quad (65)$$

The KLD between the standard multivariate normal (MVN) $p^N$ and a multivariate normal $p^N_{\mu, \Sigma} = p_{\mu, \Sigma^\frac{1}{2}}$ is

$$D_{KL}(p^N : p^N_{\mu, \Sigma}) = \frac{1}{2} \left( \text{tr}(\Sigma^{-1}) + \mu^\top \Sigma^{-1} \mu + \log |\Sigma| - d \right). \quad (66)$$
Using Corollary 3, we recover the formula for the KLD between two normal distributions with $\Sigma = \Sigma_1^{-1}\Sigma_2$ and $\mu = \Sigma_1^{-1}(\mu_2 - \mu_1)$:

$$D_{KL}(p_{\mu_1, \Sigma_1}^N : p_{\mu_2, \Sigma_2}^N) = D_{KL}\left(p : p_{\Sigma_1^{-\frac{1}{2}}(\mu_2 - \mu_1), \Sigma_1^{-1}\Sigma_2}\right),$$

$$= \frac{1}{2}\left(\text{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^\top\Sigma_2^{-1}(\mu_2 - \mu_1) + \log|\Sigma_1^{-1}\Sigma_2| - d\right). \quad (68)$$

Observe that the KLD between two multivariate normal distributions can be decomposed as the sum of a squared Mahalanobis distance

$$D_{\text{Mah}}(\mu_1, \mu_2) := \frac{1}{2}(\mu_2 - \mu_1)^\top Q(\mu_2 - \mu_1), \quad (69)$$

for $Q \succ 0$, and a scale-invariant matrix Itakura-Saito divergence

$$D_{\text{IS}}(\Sigma_1, \Sigma_2) := \frac{1}{2}\left(\text{tr}(\Sigma_2^{-1}\Sigma_1) - \log|\Sigma_2^{-1}\Sigma_1|\right),$$

also called Burg matrix divergence in [12], a matrix Bregman divergence [31]:

$$D_{KL}(p_{\mu_1, \Sigma_1}^N : p_{\mu_2, \Sigma_2}^N) = D_{\text{Mah}}^Q(\mu_1, \mu_2) + D_{\text{IS}}(\Sigma_1, \Sigma_2). \quad (71)$$

We can also derive similar results for the linear group $Y = AX + b$ of transformations for $A \in \text{GL}(d)$ (group of invertible $d \times d$ matrices) and $b \in \mathbb{R}^d$.

4 Monte Carlo estimators of $f$-divergences

Depending on the standard densities $p$ and $q$, the integrals of the $f$-divergences may be calculable in closed-form or not. When no closed-form is available, we can estimate the $f$-divergences using Monte Carlo importance sampling [35] as follows: We choose a propositional distribution $r$ and use a set $S_m = \{x_1, \ldots, x_m\} \sim_{\text{iid}} r$ of $m$ i.i.d. variates sampled from $r$ to estimate the $f$-divergence as follows:

$$\hat{I}_{f,S_m}(p : q) = \frac{1}{m} \sum_{i=1}^{m} \frac{p(x_i)}{r(x_i)} f\left(\frac{q(x_i)}{p(x_i)}\right). \quad (72)$$

In particular, when $r = p$, we end up with the following estimate often met in the literature:

$$\hat{I}_{f,S_m}(p : q) = \frac{1}{m} \sum_{i=1}^{m} f\left(\frac{q(x_i)}{p(x_i)}\right). \quad (73)$$

For example, we estimate the Kullback-Leibler divergence by $\hat{D}_{KL,S_m}(p : q) = \frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{p(x_i)}{q(x_i)}\right)$.

One of the problem of MC estimators is that they may yield inconsistent divergence measures when the proposal distribution depends on the arguments of the $f$-divergences. That is one realization (i.e., sampling with $S_m$) may find that $\hat{I}_{f,S_m}(p_1 : q) > \hat{I}_{f,S_m}(p_2 : q)$ while another realization (i.e., sampling with $S'_m$) may find that opposite result $\hat{I}_{f,S'_m}(p_1 : q) < \hat{I}_{f,S'_m}(p_2 : q)$. This lack of consistency is problematic when implementing algorithms based on divergence comparison predicates.
However, since for location-scale densities we can always reduce the calculation of \( f \)-divergences using one standard density, say:

\[
I_f(p_1, p_2 : q_1, p_2) = I_f(p : q_{P_2^{-1}(l_2-t_1)P_2^{-1}p_2}),
\]

we can estimate the \( f \)-divergences with a fixed set \( S_m \) of iid. random variates sampled from the standard density \( p \) as follows:

\[
\hat{I}_{f,S_m}(p_1, p_2) = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{q_{P_2^{-1}(l_2-t_1)P_2^{-1}p_2}(x_i)}{p(x_i)} \right)
\]

Another problem when estimating the \( f \)-divergences with Monte Carlo methods is that depending on the randomly sampled variates, we may end up with negative estimates. To overcome this problem, we shall use the following identity:

\[
I_f(p : q) = \int p(x)B_f \left( \frac{q(x)}{p(x)} : 1 \right) dx = E_p \left[ B_f \left( \frac{q(x)}{p(x)} : 1 \right) \right],
\]

where where \( B_f(a : b) \) is the scalar Bregman divergence [5]:

\[
B_f(a : b) = f(a) - f(b) - (a - b)f'(b) \geq 0.
\]

Indeed, since \( f(1) = 0 \), we have

\[
\int p(x)B_f \left( \frac{q(x)}{p(x)} : 1 \right) dx = \int p(x) \left( f \left( \frac{q(x)}{p(x)} \right) - \left( \frac{q(x)}{p(x)} - 1 \right) f'(1) \right) dx,
\]

\[
= \int p(x)f \left( \frac{q(x)}{p(x)} \right) dx - \left( \frac{q(x)}{p(x)} - 1 \right) \int_0^1 (q(x) - p(x)) dx =: I_f(p : q).
\]

Since Bregman divergences are always non-negative and equal to zero iff \( a = b \), we get another proof of Gibbs’ inequality \( I_f(p : q) \geq 0 \) (complementing the proof using Jensen’s inequality). Thus we can estimate the \( f \)-divergences non-negatively using iid. random variates \( x_1, \ldots, x_m \) from \( p(x) \) as follows:

\[
\hat{I}_f(p_1, p_2 : q_1, p_2) = \frac{1}{m} \sum_{i=1}^{m} B_f \left( \frac{q_{P_2^{-1}(l_2-t_1)P_2^{-1}p_2}(x_i)}{p(x_i)} : 1 \right) \geq 0.
\]

Furthermore, since the MC estimator of the \( f \)-divergence is the average of \( m \) scalar Bregman divergences, it follows that the estimator is a proper divergence (i.e, \( \hat{I}_f(p_1, P_1 : P_2, P_2) = 0 \Leftrightarrow (l_1, P_1) = (l_2, P_2) \)) whenever two distinct densities of the location-scale families cannot coincide in more than \( s \) points and when the random variates \( x_i \)'s have at least \( s + 1 \) distinct points.

5 Information projections onto location-scale families

We investigate how any two location-scale models \( \mathcal{F}_p \) and \( \mathcal{F}_q \) (with \( p \neq q \) and \( p, q \in \mathcal{P}(\mathcal{X}) \)) relate to each other using information projections induced by \( f \)-divergences [3, 24]. For a family of densities
Q, let \( I_f(p : Q) := \inf_{q \in Q} I_f(p : q) \) (respectively, \( I_f(P : q) := \inf_{p \in P} I_f(p : q) \)). We consider the (possibly multivariate) location-scale models as subspaces of \( \mathcal{P}(\mathbb{R}^d) \) (infinite-dimensional space) or as submodels of a multivariate location-scale model \( \mathcal{F}_m \). In the former case, we may consider nonparametric information geometry \([19, 33, 39]\) for geometrically modeling \( \mathcal{P}(\mathcal{X}) \). In the latter case, we consider the ordinary statistical manifold structure of \( \mathcal{F}_m \) (parametric information geometry \([1, 25]\)). First, let us observe that even if the KLD is asymmetric, one orientation can be finite while the reverse orientation can be infinite: For example, we have \( D_{\text{KLD}}(p^N : p^C) \simeq 0.26 < \infty \) but \( D_{\text{KLD}}(p^C : p^N) = +\infty \) where \( p^N \) denotes the standard normal density and \( p^C \) denotes the standard Cauchy density (heavy-tailed).

Recall that \( G_d = \mathbb{R}^d \times \mathbb{P}^d_+ \) denotes the \( d \)-dimensional location-scale group (or “location-positive” group). We state the remarkable projection property of a location-scale density onto another location-scale model:

**Theorem 2** (Information projection on location-scale families). The \( f \)-divergence \( I_f(p_g : q_{h^*}) = I_f(p_g : \mathcal{F}_q) \) induced by the right-sided \( f \)-divergence minimization of \( p_g \in G_d \) with \( \mathcal{F}_q \) is independent of \( g \), i.e. \( I_f(p_g : \mathcal{F}_q) = I_f(p_{g'} : \mathcal{F}_q) \) for all \( g' \in G_d \). Similarly, the \( f \)-divergence \( I_f(p_{g^*} : q_h) = I_f(\mathcal{F}_p : q_h) \) induced by the left-sided \( f \)-divergence minimization of \( q_h \) with \( \mathcal{F}_p \) is independent of \( h \), i.e. \( I_f(\mathcal{F}_p : q_h) = I_f(\mathcal{F}_p : q_{h'}) \) for all \( h' \in G_d \).

**Proof.** Using the invariance of the \( f \)-divergence under the action of \( g^{-1} \) (Theorem 1), we have

\[
\inf_{h \in G_d} I_f(p_g : q_h) = \inf_{h \in G_d} I_f(g^{-1} \circ p_g : g^{-1} \circ q_h),
\]

(82)

\[
= \inf_{\{h' = g^{-1} \circ h : h \in G_d \}} I_f(p : q_{h'}),
\]

(83)

\[
= \inf_{h' \in G_d} I_f(p : q_{h'}),
\]

(84)

since the left coset \( g^{-1}G_d \) is equal to \( G_d \). Indeed, for any \( e \in G_d \), we may find \( f \in G_d \) such that \( g^{-1}f = e \) (i.e., choose \( f = g.e \)). Let \( h^* \in G_d \) such that \( \inf_{h \in G_d} I_f(p : q_h) = I_f(p : q_{h^*}) \). Thus a minimum of \( \inf_{h \in G_d} I_f(p_g : q_h) \) is \( h^*(g) := g.h^* \) since

\[
\inf_{h \in G_d} I_f(p_g : q_h) = I_f(p : q_{h^*}) = I_f(p_g : q_{g.h^*}) = I_f(p_g : q_{h^*(g)}).
\]

(85)

Similarly, by using the conjugate generator \( f^* \), we prove that \( I_f(\mathcal{F}_p : q_h) \) is independent of \( h^* \), and a minimum of \( \inf_{g \in G_d} I_f(p_g : q_h) \) is \( g^*(h) := h.g^* \) since

\[
\inf_{g \in G_d} I_f(p_g : q_h) = I_f(p_{g^*} : q) = I_f(p_{h.g^*} : q_h) = I_f(p_{g^*(h)} : q).
\]

(86)

This property was observed without any proof in \([38]\) for the special case of the Kullback-Leibler divergence between any two univariate location-scale families. We extended this property with a proof to \( f \)-divergences between multivariate location-scale families. Notice that the projections with respect to \( f \)-divergences link orbits between the subspaces \( \mathcal{F}_p \) and \( \mathcal{F}_q \): Namely, we have the mappings \( g \mapsto h^*(g) = g.h^* \) and \( h \mapsto g^*(h) := h.g^* \).

We shall illustrate the theorem on several examples and provide some geometric interpretations of how the location-scale submodels relate to each others.
Let \( p(x) \) and \( q(x) \) be density functions defined over the support \( X = [0, \infty) \). Consider the scale families \( \mathcal{F}_p = \{ p_{s_1}(x) = \frac{1}{s_1} p\left( \frac{x}{s_1} \right) : s_1 > 0 \} \) and \( \mathcal{F}_q = \{ q_{s_2}(x) = \frac{1}{s_2} q\left( \frac{x}{s_2} \right) : s_2 > 0 \} \). Using a computer algebra system, we can show that

\[
D_{KL}(p_{s_1} : q_{s_2}) = \frac{1}{2} \left( 2 \log \frac{s_2}{s_1} + \log \frac{2}{\pi} - 1 \right) + \sqrt{\frac{2}{\pi} s_1 s_2}.
\]

Let \( r = \frac{s_1}{s_2} \). Then

\[
D_{KL}(p_{s_1} : q_{s_2}) = \sqrt{\frac{2}{\pi} r} - \log r + \log \sqrt{\frac{2}{\pi} - \frac{1}{2}}.
\]

That is, the KLD between the scale families depends only on the scale ratio as proved in Corollary 2.
Figure 2: In Euclidean geometry, parallel lines $L_1$ and $L_2$ are equidistant to each other.

The KLD is minimized wrt. to $s_2$ when $-\frac{1}{\sqrt{2}} + \sqrt{\frac{2}{\pi}} = 0$: That is, when $r = \sqrt{\frac{2}{\pi}}$ (i.e., $s_2 = s_1 \sqrt{\frac{2}{\pi}}$). We check that $D_{KL}(p_{s_1} : F_q) = \frac{1}{2} + \log \frac{2}{\pi} \simeq 0.048$ is independent of $s_1$. Thus we have $h^* = g_1^* = \sqrt{\frac{2}{\pi}}$ and $g_1^* = s_1 \sqrt{\frac{2}{\pi}}$.

Similarly, we find that $D_{KL}(F_p : q_{s_2})$ is minimized wrt $s_1$ for $s_1 = s_2$, and we have $D_{KL}(F_p : q_{s_2}) = -\frac{1}{2} \log \frac{2}{\pi} \simeq 0.226$. Figure 1 (right) illustrates geometrically the information projections between these two scale families.

Thus the location-scale densities bear some geometric similarity with parallel lines in Euclidean geometry which are equidistant as depicted in Figure 2.

Example 8. The Weibull distributions form a one-parametric family of scale families with densities expressed by:

$$p_{k,s}(x) = \frac{k}{s} \left( \frac{x}{s} \right)^{k-1} \exp \left( -\left( \frac{x}{s} \right)^k \right),$$

for $x \in X = [0, \infty)$. Parameter $s$ denotes the scale for location $l = 0$. Let $p_k(x) = p_{k,1}(x) = k x^{k-1} \exp(-x^k)$ denotes the standard density of the $k$-th Weibull scale family.

The Kullback-Leibler divergence between two Weibull distributions \cite{4} is

$$D_{KL}(p_{k_1,s_1} : p_{k_2,s_2}) = \log \frac{k_1}{s_1^{k_1}} - \log \frac{k_2}{s_2^{k_2}} + (k_1 - k_2) \left[ \log s_1 - \frac{\gamma}{k_1} \right] + \left( \frac{s_1}{s_2} \right)^{k_2} \Gamma \left( \frac{k_2}{k_1} + 1 \right) - 1. \quad (90)$$

We check that the KLD between two scale Weibull families is scale invariant:

$$\forall \lambda > 0, \quad D_{KL}(p_{k_1,\lambda s_1} : p_{k_2,\lambda s_2}) = D_{KL}(p_{k_1,s_1} : p_{k_2,s_2}), \quad (91)$$

and that

$$D_{KL}(p_{k_1,s_1} : p_{k_2,s_2}) = D_{KL}(p_{k_1,s_1} : p_{k_2,\frac{s_2}{s_1}}) = D_{KL}(p_{k_1,\frac{s_1}{s_2}} : p_{k_2}). \quad (92)$$

Indeed, we can rewrite equivalently Eq. 90 as:

$$D_{KL}(p_{k_1,s_1} : p_{k_2,s_2}) = \left( \frac{s_1}{s_2} \right)^{k_2} \Gamma \left( \frac{k_2}{k_1} + 1 \right) - k_2 \log \frac{s_1}{s_2} + \log \frac{k_1}{k_2} - \left( 1 - \frac{k_2}{k_1} \right) \gamma - 1. \quad (93)$$
This last expression highlights the use of the scale invariant ratio $\lambda = \frac{s_1}{s_2}$.

When $k_1 = k_2 = k$, the KLD between two Weibull densities of $F_{p_k}$ is:

$$D_{KL}(p_{k,s_1} : p_{k,s_2}) = \left( \frac{s_1}{s_2} \right)^k - k \log \frac{s_1}{s_2} - 1,$$

(94)
since $\Gamma(2) = 1$. In that case, since $F_{p_k}$ is an exponential family, we check that in the case the KLD amounts to the Itakura-Saito divergence (a Bregman divergence) on the swapped natural parameter $\theta_i = \frac{1}{\lambda_i}$.

The KLD between an exponential distribution ($k_1 = 1$) and a Rayleigh distribution ($k_2 = 2$) is

$$D_{KL}(p_{E,s_1} : p_{R,s_2}) = 2 \left( \frac{s_1}{s_2} \right)^2 - \log \left( \frac{s_1}{s_2} \right)^2 + c,$$

(95)

$$= 2\lambda^2 - 2 \log \lambda + c$$

(96)
since $\Gamma(2 + 1) = 2$, and where $c$ denotes a constant. It follows that $D_{KL}(p^E : p^R) = \frac{2}{\lambda^2} - \log \frac{1}{\lambda^2} + c$ is minimized for $s = \sqrt{2}$. Conversely, $D_{KL}(p^E : p^R) = 2s^2 - \log s^2 + c$ is minimized for $s = \frac{1}{\sqrt{2}}$.

The exponential and Rayleigh scale families are 1D submanifolds of the Weibull manifold whose information-geometric structure has been studied in [6].

Last but not least, these results apply for families of distributions $p_X$ that can be transformed into a location-scale family via an invertible and differentiable transformation (e.g., example 4).

A Fisher-Rao distance between two densities of a location-scale family

Let $F_p = \{ p_{l,s}(x) := \frac{1}{s} p\left( \frac{x-l}{s} \right) : (l,s) \in \mathbb{R} \times \mathbb{R}^+ \}$ be a location-scale family induced by the standard density $p(x)$ with support $\mathcal{X} = \mathbb{R}$. Location-scale families include the family of normal distributions, the family of Laplace distributions, the family of Student $t$-distributions (including the family of Cauchy distributions), the family of logistic distributions, the families of univariate elliptical distributions [22], etc.

Under mild regularity conditions (i.e., interchanging derivation and integration operation order), the Fisher information matrix (FIM) $I_\lambda(\lambda)$ with respect to parameter $\lambda = (l,s) \in \mathbb{R} \times \mathbb{R}^+$ is given by:

$$I_\lambda(\lambda) = E_{p_\lambda} \left[ \nabla_\lambda \log p_\lambda(x)(\nabla_\lambda \log p_\lambda(x))^\top \right],$$

(97)

$$= -E_{p_\lambda} \left[ \nabla_\lambda^2 \log p_\lambda(x) \right].$$

(98)

Let $g_{ij}(\lambda)$ denote the $(i,j)$-th coefficient of the FIM so that we have $I_\lambda(\lambda) = [g_{ij}(\lambda)]_{ij}$ with

$$g_{ij}(\lambda) = E_{p_\lambda} \left[ \partial_i \log p_\lambda(x) \partial_j \log p_\lambda(x) \right],$$

(99)

$$= -E_{p_\lambda} \left[ \partial_i \partial_j \log p_\lambda(x) \right],$$

(100)

where $\partial_i := \frac{\partial}{\partial x_i}$.  

17
When handling location-scale densities \( p_{l,s}(x) := \frac{1}{s} p \left( \frac{x-l}{s} \right) \), we shall observe that using a change of variable \( y = \frac{x-l}{s} \) (with \( dy = \frac{dx}{s} \)), we have for any function \( f \) the following identity:

\[
E_{p_{\lambda}} \left[ f \left( \frac{x-l}{s} \right) \right] = \int \frac{1}{s} p \left( \frac{x-l}{s} \right) f \left( \frac{x-l}{s} \right) \, dx,
\]

\[= \int p(y) f(y) dy = E_p[f(x)]. \tag{101}\]

The log-likelihood of a location-scale density is \( \log p_{l,s}(x) = \log p \left( \frac{x-l}{s} \right) - \log s \). Let us compute the coefficients of the FIM using the notations \( \partial_l = \frac{\partial}{\partial l} \) and \( \partial_s = \frac{\partial}{\partial s} \) as follows:

- Let us compute the first diagonal coefficient of the FIM using

\[
\partial_l \log p_{l,s}(x) = -\frac{1}{s} \frac{p' \left( \frac{x-l}{s} \right)}{p \left( \frac{x-l}{s} \right)}, \tag{103}\]

so that it comes that:

\[
g_{11}(\lambda) = E_{p_{\lambda}} \left[ (\partial_l \log p_{l,s}(x))^2 \right], \tag{104}\]

\[
= \frac{1}{s^2} E_{p_{\lambda}} \left[ \left( \frac{p' \left( \frac{x-l}{s} \right)}{p \left( \frac{x-l}{s} \right)} \right)^2 \right], \tag{105}\]

\[
= \frac{1}{s^2} E_p \left[ \left( \frac{p'(x)}{p(x)} \right)^2 \right]. \tag{106}\]

- We proceed and compute the second diagonal coefficient of the FIM using

\[
\partial_s \log p_{l,s}(x) = -\frac{1}{s^2} (x-l) \frac{p' \left( \frac{x-l}{s} \right)}{p \left( \frac{x-l}{s} \right)} - \frac{1}{s}, \tag{107}\]

\[
= -\frac{1}{s} \left( 1 + \frac{x-l}{s} \frac{p' \left( \frac{x-l}{s} \right)}{p \left( \frac{x-l}{s} \right)} \right), \tag{108}\]

so that it comes that

\[
g_{22}(\lambda) = E_{p_{\lambda}} \left[ (\partial_s \log p_{l,s}(x))^2 \right], \tag{109}\]

\[
= \frac{1}{s^2} E_{p_{\lambda}} \left[ \left( 1 + \frac{x-l}{s} \frac{p' \left( \frac{x-l}{s} \right)}{p \left( \frac{x-l}{s} \right)} \right)^2 \right], \tag{110}\]

\[
= \frac{1}{s^2} E_p \left[ \left( 1 + x \frac{p'(x)}{p(x)} \right)^2 \right]. \tag{111}\]

- Finally, we compute the off-diagonal coefficients of FIM as follows:

\[
g_{12}(\lambda) = g_{21} = E_{p_{\lambda}} \left[ (\partial_l \log p_{l,s}(x))(\partial_s \log p_{l,s}(x)) \right], \tag{112}\]

\[
= E_{p_{\lambda}} \left[ (\partial_l \log p_{l,s}(x))(\partial_s \log p_{l,s}(x)) \right], \tag{113}\]

18
\[ I_\lambda(\lambda) = \frac{1}{s^2} \left[ \begin{array}{cc} a^2 & c \\ c & b^2 \end{array} \right], \quad (116) \]

with the following constants depending on the standard density \( p \):

\[ a^2 = E_p \left[ \left( \frac{p'(x)}{p(x)} \right)^2 \right] \geq 0, \quad (117) \]

\[ b^2 = E_p \left[ \left( 1 + x \frac{p'(x)}{p(x)} \right)^2 \right] \geq 0, \quad (118) \]

\[ c = E_p \left[ \frac{p'(x)}{p(x)} \left( 1 + x \frac{p'(x)}{p(x)} \right) \right]. \quad (119) \]

**Proposition 3** (Fisher information of a location-scale family). The Fisher information matrix \( I(\lambda) \) of a location-scale family with continuously differentiable standard density \( p(x) \) with full support \( \mathbb{R} \) is \( I(\lambda) = \frac{1}{s^2} \left[ \begin{array}{cc} a^2 & c \\ c & b^2 \end{array} \right] \), where \( a^2 = E_p \left[ \left( \frac{p'(x)}{p(x)} \right)^2 \right] \), \( b^2 = E_p \left[ \left( 1 + x \frac{p'(x)}{p(x)} \right)^2 \right] \) and \( c = E_p \left[ \frac{p'(x)}{p(x)} \left( 1 + x \frac{p'(x)}{p(x)} \right) \right] \).

Note that when \( c \neq 0 \), the parameters \( l \) and \( s \) are correlated (i.e., non-orthogonal). Assume the standard density is an even function (e.g., the normal, Cauchy, and Laplace standard densities): We have \( p(-x) = p(x) \) and its derivative \( p'(x) \) is odd: \( p'(-x) = -p'(x) \). Then the function \( h(x) = \frac{p'(x)}{p(x)} \left( 1 + x \frac{p'(x)}{p(x)} \right) \) is odd since \( \frac{p'(x)}{p(x)} \) is odd and \( \left( 1 + x \frac{p'(x)}{p(x)} \right) \) is even. We have \( E_p[h(x)] = 0 \) for any odd function \( h(x) \) and even density \( p(x) \): Indeed, by a change of variable \( y = -x \) in the integral \( \int_{-\infty}^{0} \ldots dx \), we find that

\[ E_p[h(x)] = \int_{-\infty}^{\infty} p(x)h(x)dx, \quad (120) \]

\[ = \int_{-\infty}^{0} p(x)h(x)dx + \int_{0}^{\infty} p(x)h(x)dx, \quad (121) \]

\[ = \int_{-\infty}^{0} p(y)h(y)dy + \int_{0}^{\infty} p(x)h(x)dx, \quad (122) \]

\[ = -\int_{0}^{\infty} p(x)h(x)dx + \int_{0}^{\infty} p(x)h(x)dx, \quad (123) \]

\[ = 0. \quad (124) \]
Notice that even standard density \( p(x) \) are symmetric and have zero skewness \( E_p [x^3] \) since \( x^3 \) is an odd function.

Thus let us consider that the standard density is an even function so that the FIM with respect to parameter \( \lambda = (l, s) \) is the following diagonal matrix:

\[
I_\lambda(\lambda) = \frac{1}{s^2} \begin{bmatrix} a^2 & 0 \\ 0 & b^2 \end{bmatrix},
\]

with

\[
a^2 = E_p \left[ \left( \frac{p'(x)}{p(x)} \right)^2 \right] \geq 0,
\]

\[
b^2 = E_p \left[ \left( 1 + x \frac{p'(x)}{p(x)} \right)^2 \right] > 0.
\]

Furthermore, let us reparameterize the location-scale density by \( \theta(\lambda) = \left( \frac{b}{a} \lambda_1, \lambda_2 \right) \) where \( a = \sqrt{a^2} \) and \( b = \sqrt{b^2} \) are the positive square roots of \( a^2 \) and \( b^2 \), respectively. We have \( \lambda(\theta) = \left( \frac{b}{a} \theta_1, \theta_2 \right) \).

Using the covariance transformation of the FIM \[25\], we get

\[
I_\theta(\theta) = \left[ \frac{\partial \lambda_i}{\partial \theta_j} \right]^{\top} I_\lambda(\lambda(\theta)) \times \left[ \frac{\partial \lambda_i}{\partial \theta_j} \right]_{ij},
\]

\[
= \begin{bmatrix} \frac{b}{a} \\ 0 \end{bmatrix} \times \begin{bmatrix} a^2 \\ 0 \end{bmatrix} \times \begin{bmatrix} \frac{b}{a} \\ 0 \end{bmatrix},
\]

\[
= \frac{b^2}{\theta_2^2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
\]

This metric corresponds to a scaled metric of the Poincaré upper plane (conformal metric). It follows that the Gaussian curvature \( \kappa \) is constant and negative:

\[
\kappa = -\frac{1}{b^2} < 0.
\]

Thus the Fisher-Rao distance between two densities of a location-scale family is hyperbolic. Let \( \rho_{U,\kappa} \) denote the hyperbolic distance in the hyperbolic geometry of curvature \( \kappa \) \[29\ 30\]:

\[
\rho_{U,\kappa}(\theta_1, \theta_2) = \sqrt{-\frac{1}{\kappa} \arccosh \left( \frac{1 - \theta_1 \cdot \theta_2}{\sqrt{(1 - \theta_1 \cdot \theta_1)(1 - \theta_2 \cdot \theta_2)}} \right)},
\]

where \( \arccosh(u) = \log(u + \sqrt{u^2 - 1}) \) for \( u > 1 \) and \( \cdot \) denotes the scalar product: \( \theta \cdot \theta' = \theta^\top \theta' = \theta_1 \theta'_1 + \theta_2 \theta'_2 \).

Thus we get the following proposition:
The group SU formations) \( SU \) can be transformed into the Poincaré complex disk geometry since geodesics are straight lines restricted to the open unit disk.

**Example 10.** The Fisher-Rao distance between two Cauchy densities is a scaled hyperbolic distance between two densities \( p \).

**Proposition 4** (Fisher-Rao distance on a location-scale manifold). The Fisher-Rao distance between two densities \( p_{1,s_1} \) and \( p_{2,s_2} \) of a location-scale family \( F_p \) with even standard density \( p(x) = p(-x) \) on the support \( \mathbb{R} \) is

\[
\rho_p((l_1, s_1), (l_2, s_2)) = b \rho_U \left( \left( \frac{a}{b} l_1, s_1 \right), \left( \frac{a}{b} l_2, s_2 \right) \right),
\]

where \( a = \sqrt{E_p \left[ \left( \frac{p'(x)}{p(x)} \right)^2 \right]} \) and \( b = \sqrt{E_p \left[ \frac{p'(x) + 1}{x} \right]} > 0 \), and

\[
\rho_U((l_1, s_1), (l_2, s_2)) = \arccosh \left( \frac{1 - (l_1 l_2 + s_1 s_2)}{\sqrt{(1 - (l_1^2 + s_1^2)) \left(1 - (l_2^2 + s_2^2)\right)}} \right).
\]

**Example 9.** The Fisher-Rao distance between two normal densities \( p_{\mu_1,s_1}^{\mathcal{N}} \) and \( p_{\mu_2,s_2}^{\mathcal{N}} \) is

\[
\rho_{\mathcal{N}}((\mu_1, \sigma_1), (\mu_2, \sigma_2)) = \sqrt{2} \rho_U \left( \left( \frac{1}{\sqrt{2}} \mu_1, \sigma_1 \right), \left( \frac{1}{\sqrt{2}} \mu_2, \sigma_2 \right) \right)
\]

(133)

since \( a^2 = 1 \), \( b^2 = 2 \), \( \kappa = -\frac{1}{2} \).

**Example 10.** The Fisher-Rao distance between two Cauchy densities is a scaled hyperbolic distance

\[
\rho_{\mathcal{C}}((l_1, s_1), (l_2, s_2)) = \frac{1}{\sqrt{2}} \rho_U ((l_1, s_1), (l_2, s_2)),
\]

(134)

since \( a^2 = b^2 = \frac{1}{2} \) and \( \kappa = -\frac{1}{b^2} = -2 \).

Consider the mapping \( (l, s) \mapsto \frac{a}{b} l + is \in \mathbb{C} \) where \( i^2 = -1 \). The Poincaré complex upper plane \( \mathbb{U} \) can be transformed into the Poincaré complex disk \( \mathbb{D} \) using a Cayley transform \([29,27]\). Let \( SL_{\mathbb{R}}(2) \) be the group represented by the matrices:

\[
SL_{\mathbb{R}}(2) := \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} : a, b, c, d \in \mathbb{R}, \quad ad - bc = 1 \right\}.
\]

(135)

The action of the group \( SL_{\mathbb{R}}(2) \) on \( \mathbb{U} \) is defined by real linear fractional transforms (Möbius transformations) \( z \mapsto \frac{az + b}{cz + d} \) for \( \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL_{\mathbb{R}}(2) \) defined on the extended complex plane \( \mathbb{C} \cup \{\infty\} \). Let \( SU_{\mathbb{C}}(1,1) \) denote the special unitary group:

\[
SU_{\mathbb{C}}(1,1) := \left\{ \begin{bmatrix} a & b \\ \bar{b} & \bar{a} \end{bmatrix} : a, b \in \mathbb{C}, \quad a \bar{a} - b \bar{b} = 1 \right\}.
\]

(136)

The group \( SU_{\mathbb{C}}(1,1) \) acts on \( \mathbb{D} \) via complex linear fractional transforms: \( z \mapsto \frac{az + b}{bz + a} \). Notice that the group \( SL_{\mathbb{R}}(2) \) is isomorphic to the group \( SU_{\mathbb{C}}(1,1) \): Using the matrix representations, we have \( A \in SL_{\mathbb{R}}(2) \mapsto CAC^{-1} \in SU_{\mathbb{C}}(1,1) \) where \( C = \begin{bmatrix} 1 & -i \\ 1 & i \end{bmatrix} \). Thus we can convert \( \mathbb{U} \) to \( \mathbb{D} \) using the transformation \( \frac{z-i}{1-z} \), and reciprocally we convert \( \mathbb{D} \) to \( \mathbb{U} \) using the inverse transformation \( \frac{i(z+1)}{1-z} \). When performing geometric computing, it is preferable to use the Klein model of hyperbolic geometry since geodesics are straight lines restricted to the open unit disk.
References

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

[2] Ole Barndorff-Nielsen. *Information and exponential families*. John Wiley & Sons, 2014.

[3] Ole E Barndorff-Nielsen, Preben Blæsild, and Poul S Eriksen. *Decomposition and invariance of measures, and statistical transformation models*, volume 58. Springer Science & Business Media, 2012.

[4] Christian Bauckhage. Computing the Kullback-Leibler Divergence between two Weibull Distributions. *arXiv preprint arXiv:1310.3713*, 2013.

[5] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[6] Limei Cao, Huafei Sun, and Xiaojie Wang. The geometric structures of the Weibull distribution manifold and the generalized exponential distribution manifold. *Tamkang Journal of Mathematics*, 39(1):45–51, 2008.

[7] Edwin L Crow and Kunio Shimizu. *Lognormal distributions*. Marcel Dekker New York, 1987.

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

[9] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[10] Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

[11] Imre Csiszár and Frantisek Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490, 2003.

[12] Jason V Davis and Inderjit S Dhillon. Differential entropic clustering of multivariate Gaussians. In *Advances in Neural Information Processing Systems*, pages 337–344, 2007.

[13] Manuel Gil, Fady Alajaji, and Tamas Linder. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

[14] Masahito Hayashi. Large deviation theory for non-regular location shift family. *Annals of the Institute of Statistical Mathematics*, 63(4):689–716, 2011.

[15] Harold Hotelling. Spaces of statistical parameters. *Bull. Amer. Math. Soc*, 36:191, 1930.

[16] Robert E Kass and Paul W Vos. *Geometrical foundations of asymptotic inference*, volume 908. John Wiley & Sons, 2011.

[17] Tomo Kollo and Dietrich von Rosen. *Advanced multivariate statistics with matrices*, volume 579. Springer Science & Business Media, 2006.

[18] Fumiyasu Komaki. Bayesian prediction based on a class of shrinkage priors for location-scale models. *Annals of the Institute of Statistical Mathematics*, 59(1):135–146, 2007.

[19] John D Lafferty. The density manifold and configuration space quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.

[20] Peter D Lax. Change of variables in multiple integrals II. *The American Mathematical Monthly*, 108(2):115–119, 2001.
[21] Ann ES Mitchell. The information matrix, skewness tensor and $\alpha$-connections for the general multivariate elliptic distribution. *Annals of the Institute of Statistical Mathematics*, 41(2):289–304, 1989.

[22] Ann FS Mitchell. Statistical manifolds of univariate elliptic distributions. *International Statistical Review/Revue Internationale de Statistique*, pages 1–16, 1988.

[23] Frank Nielsen. Pattern learning and recognition on statistical manifolds: an information-geometric review. In *International Workshop on Similarity-Based Pattern Recognition*, pages 1–25. Springer, 2013.

[24] Frank Nielsen. What is... an information projection? *Notices of the AMS*, 65(3):321–324, 2018.

[25] Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020.

[26] Frank Nielsen. On Voronoi diagrams on the information-geometric Cauchy manifolds. *Entropy*, 22(7):713, 2020.

[27] Frank Nielsen. The Siegel–Klein Disk: Hilbert Geometry of the Siegel Disk Domain. *Entropy*, 22(9):1019, 2020.

[28] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[29] Frank Nielsen and Richard Nock. Hyperbolic Voronoi diagrams made easy. In *2010 International Conference on Computational Science and Its Applications*, pages 74–80. IEEE, 2010.

[30] Frank Nielsen and Richard Nock. The hyperbolic Voronoi diagram in arbitrary dimension. *arXiv preprint arXiv:1210.8234*, 2012.

[31] Richard Nock, Brice Magdalou, Eric Briys, and Frank Nielsen. Mining matrix data with Bregman matrix divergences for portfolio selection. In *Matrix Information Geometry*, pages 373–402. Springer, 2013.

[32] Atsumi Ohara and Shinto Eguchi. Geometry on positive definite matrices deformed by $V$-potentials and its submanifold structure. In *Geometric Theory of Information*, pages 31–55. Springer, 2014.

[33] Giovanni Pistone, Carlo Sempi, et al. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The annals of statistics*, 23(5):1543–1561, 1995.

[34] Yu Qiao and Nobuaki Minematsu. A study on invariance of $f$-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890, 2010.

[35] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[36] Salem Said, Lionel Bombrun, and Yannick Berthoumieu. Warped Riemannian metrics for location-scale models. In *Geometric Structures of Information*, pages 251–296. Springer, 2019.

[37] Lene Theil Skovgaard. A Riemannian geometry of the multivariate normal model. *Scandinavian journal of statistics*, pages 211–223, 1984.

[38] Cristiano Villa. A property of the Kullback–Leibler divergence for location-scale models. *arXiv preprint arXiv:1604.01983*, 2016.

[39] Jun Zhang. Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy*, 15(12):5384–5418, 2013.