On the $f$-divergences between densities of a multivariate location or scale family

Frank Nielsen  
Sony Computer Science Laboratories Inc  
Tokyo, Japan  
E-mail: Frank.Nielsen@acm.org

Kazuki Okamura  
Department of Mathematics, Faculty of Science, Shizuoka University  
Japan  
E-mail: okamura.kazuki@shizuoka.ac.jp

Abstract  
We first extend the result of Ali and Silvey [Journal of the Royal Statistical Society: Series B, 28.1 (1966), 131-142] who first reported that any $f$-divergence between two isotropic multivariate Gaussian distributions amounts to a corresponding strictly increasing scalar function of their corresponding Mahalanobis distance. We report sufficient conditions on the standard probability density function generating a multivariate location family and the function generator $f$ in order to generalize this result. This property is useful in practice as it allows to compare exactly $f$-divergences between densities of these location families via their corresponding Mahalanobis distances, even when the $f$-divergences are not available in closed-form as it is the case, for example, for the Jensen-Shannon divergence or the total variation distance between densities of a normal location family. Second, we consider $f$-divergences between densities of multivariate scale families: We recall Ali and Silvey’s result that for normal scale families we get matrix spectral divergences, and we extend this result to densities of a scale family.

Keywords: $f$-divergence; Jensen-Shannon divergence; multivariate location-scale family; spherical distribution; affine group; multivariate normal distributions; multivariate Cauchy distributions; matrix spectral divergence.

1 Introduction

Let $\mathbb{R}$ denote the real field and $\mathbb{R}_{++}$ the set of positive reals. The $f$-divergence $[6][1]$ induced by a convex generator $f : \mathbb{R}_{++} \to \mathbb{R}$ between two probability density functions (PDFs) $p(x)$ and $q(x)$ defined on the full support $\mathbb{R}^d$ is defined by

$$I_f(p : q) := \int p(x) f \left( \frac{q(x)}{p(x)} \right) \, dx.$$  

It follows from Jensen’s inequality that we have

$$I_f(p : q) \geq f \left( \int p(x) \frac{q(x)}{p(x)} \, dx \right) = f(1).$$
| $f$-divergence               | Formula $I_f(p : q)$                                      | Generator $f(u)$ with $f(1) = 0$ |
|----------------------------|----------------------------------------------------------|----------------------------------|
| Total variation            | $\frac{1}{2} \int |p(x) - q(x)| dx$                                      | $\frac{1}{2}|u - 1|$            |
| Squared Hellinger          | $\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx$                | $(\sqrt{u} - 1)^2$              |
| Pearson $\chi^2$          | $\int \frac{(p(x) - p(x))^2}{p(x)} dx$                | $(u - 1)^2$                     |
| Neyman $\chi^2$           | $\int \frac{(p(x) - q(x))^2}{q(x)} dx$                | $\frac{(1-u)^2}{u}$             |
| Kullback-Leibler           | $\int p(x) \log \frac{p(x)}{q(x)} dx$                | $- \log u$                      |
| reverse Kullback-Leibler   | $\int q(x) \log \frac{q(x)}{p(x)} dx$                | $u \log u$                      |
| Jeffreys                   | $\int (p(x) - q(x)) \log \frac{p(x)}{q(x)} dx$       | $(u - 1) \log u$                |
| Jensen-Shannon             | $h \left( \frac{p+q}{2} \right) - \frac{h(p)+h(q)}{2}$ | $-(u + 1) \log \frac{1+u}{2} + u \log u$ |
|                           | where $h(p) = \int p(x) \log \frac{1}{p(x)} dx$ is Shannon entropy |

Table 1: Some common statistical divergences expressed as $f$-divergences.

Thus we shall consider convex generators $f(u)$ such that $f(1) = 0$. Moreover, in order to ensure that $I_f(p : q) = 0$ if and only if $p(x) = q(x)$ except at countably many points $x$, we need $f(u)$ to be strictly convex at $u = 1$. The class of $f$-divergences include the total variation distance (the only $f$-divergence up to scaling which is a metric distance [15]), the Kullback-Leibler divergence (and its two common symmetrizations, namely, the Jeffreys divergence and the Jensen-Shannon divergence), the squared Hellinger divergence, the Pearson and Neyman $\chi^2$-divergences, etc. The formula for those statistical divergences with their corresponding generators are listed in Table 1.

Let $N(\mu, \Sigma) \sim p_{\mu, \Sigma}(x)$ be a multivariate normal (MVN) distribution with mean $\mu \in \mathbb{R}^d$ and positive-definite covariance matrix $\Sigma \in \text{Sym}_{++}(d)$ (where $\text{Sym}_{++}(d)$ denotes the set of positive-definite matrices), where the PDF is defined by

$$p_{\mu, \Sigma}(x) = \frac{1}{\sqrt{\det(2\pi \Sigma)}} \exp \left( -\frac{1}{2} (x - \mu) \Sigma^{-1} (x - \mu) \right).$$

In their landmark paper, Ali and Silvey [11] (Section 6 of [11], pp. 141-142) mentioned the following two properties of $f$-divergences between MVN distributions:

P1. The $f$-divergences $I_f(p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma})$ between two MVN distributions $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ with prescribed covariance matrix $\Sigma$ is an increasing function of their Mahalanobis distance [15] $\Delta_\Sigma(\mu_1, \mu_2)$, where

$$\Delta_\Sigma^2(\mu_1, \mu_2) := (\mu_2 - \mu_1) \Sigma^{-1} (\mu_2 - \mu_1).$$

Ali and Silvey briefly sketched a proof by considering the following property obtained from a change of variable $y = \frac{(x-\mu_1) \Sigma^{-1} (\mu_2 - \mu_1)}{\Delta_\Sigma(\mu_1, \mu_2)} \in \mathbb{R}$:

$$I_f(p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) = I_f(p_{0,1} : p_{\Delta_\Sigma(\mu_1, \mu_2), 1}).$$
That is, the $f$-divergences between multivariate normal distributions with prescribed covariance matrix (left-hand side) amount to corresponding $f$-divergences between the univariate normal distribution $N(0, 1)$ and $N(\Delta_\Sigma(\mu_1, \mu_2), 1)$ (right-hand side).

P2. The $f$-divergences $I_f(p_{\mu_1, \Sigma_1} : p_{\mu_2, \Sigma_2})$ between two $\mu$-centered MVN distributions $N(\mu, \Sigma_1)$ and $N(\mu, \Sigma_2)$ is an increasing function of the terms $|1 - \lambda_i|^s$, where the $\lambda_i$'s denote the eigenvalues of matrix $\Sigma_2\Sigma_1^{-1}$. That is, the $f$-divergences between $\mu$-centered MVN distributions are spectral matrix divergences [13].

In this paper, we investigate whether these two properties hold or not for multivariate location families and multivariate scale families which generalize the multivariate centered (same mean) normal families $N_{\mu} := \{N(\mu, \Sigma) : \Sigma \in \text{Sym}_{++}(d)\}$ and the multivariate isotropic (same covariance) normal distributions $N_{\Sigma} := \{N(\mu, \Sigma) : \mu \in \mathbb{R}^d\}$, respectively.

We summarize our main contributions as follows:

- We extend property P1 to arbitrary multivariate location families in Theorem 1 under Assumption 1 (i.e., spherical distribution of the standard PDF with $f(u) \in C^2$).
- We illustrate property P1 for the multivariate location normal distributions and the multivariate location Cauchy distributions for various $f$-divergences, and discuss practical computational applications in Section 3.3.
- We then report the spectral matrix $f$-divergences (Property P2) for the multivariate scale normal distributions for the Kullback-Leibler divergence and the $\alpha$-divergences in Section 4, and generalize this property to densities of a scale family (Proposition 3).

The paper is organized as follows: We first describe generic multivariate location-scale families including the action of the affine group on $f$-divergences in Section 2. In Section 3, we present our main theorem which generalizes property P1, and illustrate the theorem with examples of $f$-divergences between multivariate location normal or location Cauchy families. Finally, we show that the $f$-divergences between densities of a scale family is always a spectral matrix divergence (Proposition 3).

2 The $f$-divergences between multivariate location-scale families

2.1 Multivariate location-scale families

Let $p(x)$ be any arbitrary PDF defined on the full support $\mathbb{R}^d$. The multivariate location-scale family $\mathcal{P}$ is defined as the set of distributions with PDFs:

$$p_{l,P}(x) := \det(P)^{-1} p \left( P^{-1} (x - l) \right), \quad x \in \mathbb{R}^d,$$

where the set of multivariate location-scale parameters $(l, P)$ belongs to $\mathbb{H}_d := \mathbb{R}^d \times \text{Sym}_{++}(d)$. PDF $p(x)$ is called the standard density since $p(x) = p_{0,I}(x)$ where $I$ denote the $d$-dimensional identity matrix.

By considering $P = (\Sigma^{1/2})^2$ where $\Sigma^{1/2}$ denotes the unique square root of the covariance matrix $\Sigma$, and letting $l = \mu$, we can express the PDFs of $\mathcal{P}$ as

$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{\det(\Sigma)}} p \left( \Sigma^{-1/2} (x - \mu) \right), \quad x \in \mathbb{R}^d. \quad (1)$$
The multivariate location-scale families generalize the univariate location-scale families with \( \Sigma = \sigma^2 \):
\[
P = \left\{ \frac{1}{\sigma} p \left( \frac{x - \mu}{\sigma} \right) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_{++} \right\}.
\]

In the reminder, we shall focus on the following two multivariate location-scale families:

MVN. MultiVariate Normal location families:
\[
p(x) = \frac{1}{(2\pi)^{d/2}} \exp \left( -\frac{x^2}{2} \right), \quad x \geq 0.
\]
Notice that for \( X \sim N(0, I_d) \), we have \( Y = PX + l \sim N(l, PP^T) \). Thus when \( l = \mu \) and \( P = \Sigma^{1/2} \), we get \( Y \sim N(\mu, \Sigma) \).

MVS. MultiVariate Student location families with \( \nu \) degree(s) of freedom:
\[
p(x) = \frac{\Gamma((\nu + d)/2)}{\Gamma(\nu/2)(\nu\pi)^{d/2}} \frac{1}{(1 + x/\nu)^{(\nu + d)/2}}, \quad x \geq 0,
\]
where \( \Gamma(t) \) is the Gamma function:
\[
\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx.
\]
The case of \( \nu = 1 \) corresponds to the MultiVariate Cauchy (MVC) family. The probability density function of a MVC is
\[
p_{\mu,\Sigma}(x) = \frac{\Gamma \left( \frac{d+1}{2} \right)}{\pi^{d/2} \det(\Sigma)^{1/2} (1 + \Delta_{\Sigma}(x, \mu))^{(d+1)/2}}.
\]
Let us note in passing that if random vector \( X = (X_1, \ldots, X_d) \) follows the standard MVC, then \( X_1, \ldots, X_d \) are not statistically independent \[17\]. Thus the MVC family differs from the MVN family from that viewpoint.

Multivariate location-scale families also include the multivariate elliptical distributions \[18\] \[12\].

### 2.2 Action of the affine group

The family \( \mathcal{P} = \{ p_{l, P} : (l, P) \in \mathbb{H} \} \) can also be obtained by the action (denoted by the dot .) of the affine group \( \text{Aff}(\mathbb{R}^d) := \mathbb{R}^d \rtimes \text{GL}_d(\mathbb{R}) \) \[10\] (where \( \text{GL}_d(\mathbb{R}) \) denotes the General Linear \( d \)-dimensional group) on the standard PDF: \( \mathcal{P} = \{ p_{l, P}(x) = (l, P).p(x), (l, P) \in \text{Aff}(\mathbb{R}^d) \} \), where the group is equipped with the following semidirect product:
\[
(l_1, A_1).(l_2, A_2) := (l_1 + A_1l_2, A_1A_2).
\]
The inverse element of \( (l, A) \) is \( (-A^{-1}l, A^{-1}) \). One can check that \( (l, A).(-A^{-1}l, A^{-1}) = (l, A) = (0, I_d) \), where \( I \) denotes the \( d \times d \) identity matrix.

The affine group \( \text{Aff}(\mathbb{R}^d) \) can be handled as a matrix group by mapping its elements \( (l, A) \) to corresponding \( (d + 1) \times (d + 1) \) matrices as follows:
\[
(l, A) \leftrightarrow \begin{bmatrix} A & l \\ 0 & 1 \end{bmatrix}.
\]
That is, \( \text{Aff}(\mathbb{R}^d) \) can be interpreted as a subgroup of \( \text{GL}(d + 1) \).
2.3 Affine group action on $f$-divergences

The $f$-divergences between two PDFs $p_{l_1} P_{l_1}$ and $p_{l_2} P_{l_2}$ of a multivariate location-scale family $P$ are invariant under the action of the affine group [18]:

$$I_f ((l, P) . p_{l_1} P_{l_1} : (l, P) . p_{l_2} P_{l_2}) = I_f (p_{l_1} P_{l_1} : p_{l_2} P_{l_2}).$$

Thus by choosing the inverse element $(-P_1^{-1} l_1, P_1^{-1})$ of $(l_1, P_1)$, we get the following Proposition [18]:

**Proposition 1.** We have

$$I_f (p_{l_1} P_{l_1} : p_{l_2} P_{l_2}) = I_f ((-P_1^{-1} l_1, P_1^{-1}), p_{l_1} P_{l_1} : (-P_1^{-1} l_1, P_1^{-1}), p_{l_2} P_{l_2}),$$

$$= I_f (p_0 P_{l_0} : P_{l_0}^{-1} (l_2 - l_1), P_{l_0}^{-1} P_{l_2}).$$

**Remark 1.** When $f(u) = - \log u$, we get the Kullback-Leiber divergence and we have:

$$D_{KL} (p_{l_1} P_{l_1} : p_{l_2} P_{l_2}) = D_{KL} (p_0 P_{l_0} : P_{l_0}^{-1} (l_2 - l_1), P_{l_0}^{-1} P_{l_2}).$$

Since the KLD is the difference of the cross-entropy minus the entropy [5] (hence also called relative entropy), we have

$$D_{KL} (p_{l_1} P_{l_1} : p_{l_2} P_{l_2}) = h \left( p_0 P_{l_0} : P_{l_0}^{-1} (l_2 - l_1), P_{l_0}^{-1} P_{l_2} \right) - h(p_0 P_{l_0}),$$

where $h(p : q) = - \int p(x) \log q(x) dx$ is the cross-entropy between $p(x)$ and $q(x)$, and $h(p) = h(p : q)$ is the differentiable entropy. When both $p(x) = p_{\mu_1, \Sigma_1}(x)$ and $q(x) = p_{\mu_2, \Sigma_2}(x)$ are $d$-variate normal distributions, the KLD can be decomposed as the sum a squared Mahalanobis distance $\Delta_{\Sigma_2^{-1}}(\mu_1, \mu_2)$ plus a matrix Burg divergence [7] $D_B(\Sigma_1, \Sigma_2)$:

$$D_{KL}(p_{\mu_1, \Sigma_1} : p_{\mu_2, \Sigma_2}) = \frac{1}{2} D_B(\Sigma_1, \Sigma_2) + \frac{1}{2} \Delta_{\Sigma_2^{-1}}(\mu_1, \mu_2),$$

where the matrix Burg divergence is defined by

$$D_B(\Sigma_1, \Sigma_2) = \text{tr}(\Sigma_2 \Sigma_1^{-1}) + \log \det (\Sigma_2 \Sigma_1^{-1}) - d.$$

However, the KLD between two Cauchy distributions (viewed like normal distributions as a location-scale family) cannot be decomposed as the sum a squared Mahalanobis distance and another divergence [24].

In particular, we have for PDFs with the same scale matrix $P$, the following identity:

$$I_f (p_{l_1} P_{l_1} : p_{l_2} P_{l_2}) = I_f \left( p_0 P_{l_0} : P_{l_0} (l_2 - l_1) \right) = I_f \left( p P_{(l_1 - l_2)} : p_0 P_{l_0} \right).$$

Let $\Sigma := PP^\top$ so that $P = \Sigma^{\frac{1}{2}}$. The mapping $P \leftrightarrow \Sigma^{\frac{1}{2}}$ is a diffeomorphism on the open cone $\text{Sym}_+ (d)$ of positive-definite matrices.

Thus we have

$$I_f (p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) = I_f \left( p_0 P_{l_0} : \Sigma^{-\frac{1}{2}} (\mu_2 - \mu_1) \right) = I_f \left( p_0 P_{l_0} : \Sigma^{-\frac{1}{2}} (\mu_1 - \mu_2) \right).$$

Note that in general, we have $\mu \neq E_{X \sim p_{\mu, \Sigma}[X]}$ and $\Sigma \neq \text{Cov}_{X \sim p_{\mu, \Sigma}}[X]$. However, for the special case of multivariate normal family, we have both $\mu = E_{X \sim p_{\mu, \Sigma}[X]}$ and $\Sigma = \text{Cov}_{X \sim p_{\mu, \Sigma}}[X]$. 

5
3 The \(f\)-divergences between densities of a multivariate location family

Let us define the squared Mahalanobis distance \([15]\) between two MVNs \(N(\mu_1, \Sigma)\) and \(N(\mu_2, \Sigma)\) as follows:

\[
\Delta^2_\Sigma(\mu_1, \mu_2) := (\mu_2 - \mu_1)^\top \Sigma^{-1}(\mu_2 - \mu_1).
\]

Since the covariance matrix \(\Sigma\) is positive-definite, we have \(\Delta^2_\Sigma(\mu_1, \mu_2) \geq 0\) and zero if and only if \(\mu_1 = \mu_2\). The squared Mahalanobis distance generalizes the squared Euclidean distance when \(\Sigma = I\): \(\Delta^2_I(\mu_1, \mu_2) = \|\mu_1 - \mu_2\|^2\).

The Kullback-Leibler divergence \([5]\) between two multivariate isotropic normal distributions follows:

\[
D_{\text{KL}}(p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) = \int_{x \in \mathbb{R}^d} p_{\mu_1, \Sigma}(x) \log \frac{p_{\mu_1, \Sigma}(x)}{p_{\mu_2, \Sigma}(x)} \, dx = \frac{1}{2} \Delta^2_\Sigma(\mu_1, \mu_2),
\]

where

\[
D_{\text{KL}}(p : q) = \int p(x) \log \left(\frac{p(x)}{q(x)}\right) \, d\mu(x) = I_{\text{KL}}(p : q)
\]

for \(f_{\text{KL}}(u) = -\log u\).

Moreover, the PDD of a MVN with covariance matrix \(\Sigma\) and mean \(\mu\) can also be written using the squared Mahalanobis distance as follows:

\[
p(x; \mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2} \Delta^2_\Sigma(x, \mu)\right).
\]

This rewriting highlights the general duality between Bregman divergences (e.g., the squared Mahalanobis distance is a Bregman divergence) and the exponential families \([3]\) (e.g., multivariate Gaussian distributions).

We shall make the following set of assumptions for the standard density \(p(x)\) and \(f(u)\):

**Assumption 1.** (i) We assume that there exists a function \(p : [0, \infty) \rightarrow (0, \infty)\) such that \(p\) is in \(C^1\) class, \(p'(x) < 0, x \geq 0\), and furthermore \(p(x) = \tilde{p}(||x||^2)\), \(x \in \mathbb{R}^d\).

(ii) We assume that \(f : (0, \infty) \rightarrow \mathbb{R}\) satisfies that it is in \(C^2\) class, \(f(1) = 0\) and \(f''(x) > 0, x > 0\).

(iii) For every \(t \in \mathbb{R}^d\),

\[
\int_{\mathbb{R}^d} \left| f' \left( \frac{\tilde{p}(||x + t||^2)}{\tilde{p}(||x||^2)} \right) \right| \, d\tilde{p}(||x||^2) < +\infty.
\]

(iv) For every compact subset \(K\) of \(\mathbb{R}^d\),

\[
\int_{\mathbb{R}^d} \sup_{t \in K} \left| f' \left( \frac{\tilde{p}(||y + t||^2)}{\tilde{p}(||y||^2)} \right) \right| \, d\tilde{p}(||y||^2) \|y\| \, dy < +\infty.
\]

Notice that using the assumption that the standard density \(p(x) = \tilde{p}(||x||^2)\), we get

\[
p_{\mu, \Sigma}(x) = \frac{1}{\sqrt{\det(\Sigma)}} \tilde{p} \left( \Sigma^{-1/2}(x - \mu) \right),
\]

\[
= \frac{1}{\sqrt{\det(\Sigma)}} \tilde{p}(\Delta^2_\Sigma(x, \mu)),
\]

\[
= \frac{1}{\sqrt{\det(\Sigma)}} \tilde{p}(\Delta^2_\Sigma(x, \mu)),
\]
since $|\Sigma^{-\frac{1}{2}}(x - \mu)|^2 = (\Sigma^{-\frac{1}{2}}(x - \mu))^T \Sigma^{-\frac{1}{2}}(x - \mu) = \Delta^2_{\Sigma}(x, \mu)$. Thus the standard density $p(x)$ of the multivariate location-scale family $\mathcal{P}$ is the density of a spherical distribution \cite{25} and $p_{\mu, \Sigma}(x)$ is the density of an elliptical symmetric distribution \cite{22}.

We let

$$I_f (p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) := \int_{\mathbb{R}^d} f\left( \frac{p_{\mu_2, \Sigma}(x)}{p_{\mu_1, \Sigma}(x)} \right) p_{\mu_1, \Sigma}(x)dx.$$ 

This is well-defined due to Assumption [1 (iii).

We state the main theorem generalizing [1:

**Theorem 1** (f-divergence between location families). Under Assumption [1 there exists a strictly increasing and differentiable function $h_f$ such that

$$I_f (p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) = h_f \left( \Delta^2_{\Sigma}(\mu_1, \mu_2) \right), \quad \mu_1, \mu_2 \in \mathbb{R}^d. \tag{2}$$

**Proof.** **Step 1.** We let $d = 1$ and $\Sigma = 1$.

$$I_f (p_{\mu_1,1} : p_{\mu_2,1}) = I_f (p_{0,0} : p_{\mu_2 - \mu_1,1}) = \int_{\mathbb{R}} f\left( \frac{\tilde{p}(x - (\mu_2 - \mu_1))^2}{\tilde{p}(x^2)} \right) \tilde{p}(x^2)dx.$$ 

Let

$$F(t) := I_f (p_{0,1} : p_{t,1}) = \int_{\mathbb{R}} f\left( \frac{\tilde{p}(x - t)^2}{\tilde{p}(x^2)} \right) \tilde{p}(x^2)dx, \quad t \in \mathbb{R}.$$ 

By Assumption [1 (i), $F(t) = F(-t), \quad t \in \mathbb{R}$. Hence if we let $h_f(s) := F(\sqrt{s})$, then $h_f$ satisfies Eq. (2).

Now it suffices to show that $F$ is strictly increasing and differentiable. By Assumption [1 (iv),

$$F'(t) = \int_{\mathbb{R}} \frac{d}{dt} \left( f\left( \frac{\tilde{p}(x - t)^2}{\tilde{p}(x^2)} \right) \right) \tilde{p}(x^2)dx$$ 

$$= 2 \int_{\mathbb{R}} f' \left( \frac{\tilde{p}(x - t)^2}{\tilde{p}(x^2)} \right) (t - x)\tilde{p}'((x - t)^2)dx$$ 

$$= 2 \int_{\mathbb{R}} f' \left( \frac{\tilde{p}(y^2)}{\tilde{p}((y + t)^2)} \right) y\tilde{p}'(y^2)dy$$ 

$$= -2 \int_{0}^{\infty} \left( f' \left( \frac{\tilde{p}(y^2)}{\tilde{p}((y + t)^2)} \right) - f' \left( \frac{\tilde{p}(y^2)}{\tilde{p}((y - t)^2)} \right) \right) y\tilde{p}'(y^2)dy.$$ 

By the mean-value theorem and Assumption [1 (i) and (ii),

$$f' \left( \frac{\tilde{p}(y^2)}{\tilde{p}((y + t)^2)} \right) - f' \left( \frac{\tilde{p}(y^2)}{\tilde{p}((y - t)^2)} \right) > 0, \quad \tilde{p}'(y^2) < 0,$$

for every $y, t > 0$. Hence $F'(t) > 0$ for every $t > 0$.

**Step 2.** We let $\Sigma = I$.

$$I_f (p_{\mu_1, I_d} : p_{\mu_2, I_d}) = I_f (p_{0, I} : p_{\mu_2 - \mu_1, I_d}) = \int_{\mathbb{R}^d} f\left( \frac{\tilde{p}(\|x - (\mu_2 - \mu_1)\|^2)}{\tilde{p}(\|x\|^2)} \right) \tilde{p}(\|x\|^2)dx.$$
Let
\[ F(t) := I_f(p_{0,1} : p_{t,T}) = \int_{\mathbb{R}^d} f \left( \frac{\tilde{p}(\|x-t\|^2)}{\tilde{p}(\|x\|^2)} \right) \tilde{p}(\|x\|^2) dx, \quad t \in \mathbb{R}^d. \] (3)

By changing the variable \( x \) by an orthogonal matrix, \( F(s) = F(t) \) if \( \|s\| = \|t\| \). Hence we can assume that \( t = (t, 0, \ldots, 0)^T \in \mathbb{R}^d, t > 0 \). For \( x_2, \ldots, x_d \in \mathbb{R} \), let
\[ F_{x_2,\ldots,x_d}(t) := \int_{\mathbb{R}} f \left( \frac{\tilde{p}(\|x_1-t\|^2 + x_2^2 + \cdots + x_d^2)}{\tilde{p}(\|x\|^2)} \right) \tilde{p}(\|x\|^2) dx_1, \quad t \in \mathbb{R}. \]

Then, we can show that \( F'_{x_2,\ldots,x_d}(t) > 0 \) for every \( t > 0 \) and every \( x_2, \ldots, x_d \in \mathbb{R} \), in the same manner as in Step 1. By this and Assumption 1 (iv),
\[ F'(t) = \int_{\mathbb{R}^{d-1}} F'_{x_2,\ldots,x_d}(t) dx_2 \cdots dx_d > 0, \quad t > 0. \]

Hence if we let \( h_f(s) := F(\sqrt{s}) \), then \( h_f \) satisfies Eq. (2).

**Step 3.** Finally, we consider the general case. Let \( \mu := \Sigma^{-1/2}(\mu_2 - \mu_1) \). Then,
\[ I_f(p_{\mu_1,\Sigma} : p_{\mu_2,\Sigma}) = \sqrt{\det\Sigma} I_f(p_{0,I_d} : p_{\mu,I_d}). \]

Hence this case is attributed to Step 2.

We shall now illustrate this theorem with several examples.

### 3.1 The normal location families

#### 3.1.1 The scalar functions \( h_f \) for some common \( f \)-divergences

Table 2 lists some examples of \( f \)-divergences with their corresponding monotone increasing functions \( h_f \). We consider the following \( f \)-divergences between two PDFs of MVNs with same covariance matrix:

- For the \( \chi^2 \) divergence with \( f_{\chi^2}(u) = (u-1)^2 \), we have \( h_{\chi^2}(u) = h_{\chi^2}(u) = 1 - \exp(-\frac{1}{2}u) \), and more generally for the order-\( k \) chi divergences (\( f \)-divergence generator \( f_{\chi,k}(u) = (u-1)^k \)) between \( p_{\mu_1,\Sigma} \) and \( p_{\mu_2,\Sigma} \), we get [20]:

\[ h_{\chi,k}(u) = \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} \exp \left( \frac{1}{2} i(i-1) \Delta^2_{\Sigma}(\mu_1, \mu_2) \right). \]

We observe that the order-\( k \) \( \chi \)-divergence between isotropic MVNs diverges as \( k \) increases. It is easy to check that we can compute \( h_f \) in closed form for any convex polynomial \( f \)-divergence generator \( f(u) \).

- For the Kullback-Leibler divergence, we have \( D_{KL}[p_{\mu_1,\Sigma} : p_{\mu_2,\Sigma}] = \frac{1}{2} \Delta^2_{\Sigma}(\mu_1, \mu_2) \) so \( h_{KL}(u) = \frac{1}{2} u \). Notice that because \( f \)-divergences between isotropic MVNs are symmetric, the Chernoff information coincides with the Bhattacharyya distance.
\[
\begin{array}{|l|l|
\hline
f\text{-divergence} & f(u) \text{ and } h_f(u) \\
\hline
\chi^2\text{-squared divergence} & (u - 1)^2 \text{ and } 1 - \exp\left(-\frac{1}{2}u\right) \\
\text{Order-}k \ \chi \ \text{divergence} & (u - 1)^k \text{ and } \sum_{i=0}^{k}(-1)^{k-i}(k) \exp\left(\frac{1}{2}i(i - 1)u\right) \\
\text{Kullback-Leibler divergence} & -\log(u) \text{ and } \frac{1}{4}u \\
squared \text{Hellinger divergence} & (\sqrt{u} - 1)^2 \text{ and } 1 - \exp\left(-\frac{1}{4}u\right) \\
\text{Amari’s } \alpha\text{-divergence} & \frac{4}{1 - \alpha^2} \left(1 - u^{\frac{1+\alpha}{2}}\right) \text{ and } \frac{4}{1 - \alpha^2} \left(1 - \exp\left(-\frac{1-\alpha^2}{8}u\right)\right) \\
\text{Jensen-Shannon divergence} & u \log u - (1 + u) \log \frac{1+u}{2} \text{ and } \frac{1}{4}u - I_{JS}(u) \\
\text{Total variation distance} & |u - 1| \text{ and } 1 - 2Q\left(\frac{1}{\sqrt{2}}\sqrt{u}\right) := 1 - 2 \int_{\frac{1}{\sqrt{2}}\sqrt{u}}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}t^2)\,dt \\
\hline
\end{array}
\]

Table 2: The \( f \)-divergences between two normal distributions with identical covariance matrix can always be expressed as an increasing function of the squared Mahalanobis distance: \( I_f(p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) = h_f(\Delta^2_{\Sigma}(\mu_1, \mu_2)) \).

- The total variation divergence (a metric \( f \)-divergence obtained for \( f_{TV}(u) = |u - 1| \) which is always upper bounded by \( 1 \)) between two multivariate Gaussians \( p_{\mu_1, \Sigma} \) and \( p_{\mu_2, \Sigma} \) with the same covariance matrix is reported indirectly in [24]: The probability of error \( P_e \) (with \( P_e \leq \frac{1}{2} \)) in Bayesian binary hypothesis with equal prior is \( P_e(p_1, p_2) = \frac{1}{2}(1 - D_{TV}(p_1, p_2)) = Q\left(\frac{1}{2}\|\Sigma^{-\frac{1}{2}}(\mu_2 - \mu_1)\|\right) \) (Eq. (2) in [24]) where \( Q(x) = 1 - Q(-x) = 1 - \Phi(x) \) where \( \Phi(x) \) denotes the cumulative distribution function of the standard normal distribution. So we get the function \( h_{TV} \) as a definite integral of a function of a squared Mahalanobis distance:

\[
D_{TV}[p_1, p_2] = 1 - 2P_e(p_1, p_2) = 1 - 2Q\left(\frac{1}{2}\sqrt{\Delta^2_{\Sigma}(\mu_1, \mu_2)}\right).
\]

**Remark 2.** Notice that the Fisher-Rao distance between two multivariate normal distributions with the same covariance matrix is also a monotonic increasing function of their Mahalanobis distance [3]:

\[
\rho(p_{\mu_1, \Sigma}, p_{\mu_2, \Sigma}) = \sqrt{2} \arccosh \left(1 + \frac{\Delta^2_{\Sigma}(\mu_1, \mu_2)}{4}\right),
\]

where \( \arccosh(x) = \log(x + \sqrt{x^2 - 1}) \) for \( x \geq 1 \). That is, we have \( h_\rho(u) = \sqrt{2} \arccosh (1 + \frac{u}{4}) \).

### 3.1.2 The special case of the Jensen-Shannon divergence

The Jensen-Shannon divergence [11] is a symmetrization of the Kullback-Leibler divergence:

\[
D_{JS}[p, q] = \frac{1}{2} \left( D_{KL} \left[ p : \frac{p + q}{2} \right] + D_{KL} \left[ q : \frac{p + q}{2} \right] \right).
\]

The JSD is a \( f \)-divergence for the generator \( f_{JS}(u) = u \log u - (1 + u) \log \frac{1+u}{2} \), is always upper bounded by \( \log 2 \), and can further be embedded into a Hilbert space [9]. The JSD can be interpreted in information theory as the transmission rate in a discrete memoryless channel [9].
Although we do not have a closed-form formula for the Jensen-Shannon divergence $D_{JS}[p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}]$, knowing that $D_{JS}[p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}] = h_{f_{JS}}(\Delta^2_{\Sigma}(\mu_1, \mu_2))$, allows one to compare exactly the JSDs since $f_{JS}$ is a strictly increasing function. That is, we have the equivalence of following signs of the predicates:

$$D_{JS}[p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}] > D_{JS}[p_{\mu_3, \Sigma} : p_{\mu_4, \Sigma}] \iff \Delta^2_{\Sigma}(\mu_1, \mu_2) > \Delta^2_{\Sigma}(\mu_3, \mu_4).$$

In [16], a formula for the differential entropy of the Gaussian mixture $m(x; \mu, \sigma) = \frac{1}{2}p_{-\mu, \sigma}(x) + \frac{1}{2}p_{-\mu, \sigma}(x)$ is reported using a definite integral which we translate using the squared Mahalanobis distance as follows:

$$h(m(x; \mu, \sigma)) = \frac{1}{2} \log(2\pi e\sigma^2) + \frac{1}{4}\Delta_{\sigma^2}(-\mu, \mu) - I_{JS}(\Delta_{\sigma^2}(-\mu, \mu)),$$

where

$$I_{JS}(\Delta^2) := \sqrt{\frac{8}{\pi\Delta^2}} \exp\left(-\frac{\Delta^2}{8}\right) \int_0^\infty e^{-\frac{2x}{\Delta^2}} \cosh(x) \log \cosh(x) dx.$$  

We have $I_{JS}(0) = 0$ and the function $I_{JS}$ can be tabulated as in [16].

Since the Jensen-Shannon divergence between two distributions amounts to the differential entropy of the mixture minus the average of the mixture entropies, we get

$$D_{JS}[p_{\mu_1, \Sigma}, p_{\mu_2, \Sigma}] = D_{JS}[p_{0,1, p\Delta_{\Sigma}(\mu_1, \mu_2), 1}]$$

$$= \frac{1}{2} \log \det(2\pi e\Sigma) + \frac{1}{4}\Delta_{\Sigma^2}(\mu_1, \mu_2) - I_{JS}(\Delta_{\Sigma^2}(\mu_1, \mu_2)) - \frac{1}{2} \log \det(2\pi e\Sigma)$$

$$= \frac{1}{4}\Delta_{\Sigma^2}(\mu_1, \mu_2) - I_{JS}(\Delta_{\Sigma^2}(\mu_1, \mu_2)),$$

since $h[p_{\mu_1, \Sigma}] = h[p_{\mu_2, \Sigma}] = \frac{1}{2} \log \det(2\pi e\Sigma) = \frac{d}{2} \log(2\pi e) + \frac{1}{2} \log |\Sigma|$.  

In Section 3.3, we graph the functions $h_f$ for the total variation distance and the Jensen-Shannon divergence.

### 3.2 Cauchy location family

Notice that for $d = 1$, since the $\chi^2$-divergence (a f-divergence for $f(u) = (u - 1)^2$) between two Cauchy location densities $p_{l_1, s}$ and $p_{l_2, s}$ with prescribed scale $s$ is [21]:

$$D_{\chi^2}(p_{l_1, s} : p_{l_2, s}) = \frac{(l_2 - l_1)^2}{2s^2} \equiv \chi_s(l_1, l_2),$$

we have

$$D_{\chi^2}(p_{l_1, s} : p_{l_2, s}) = \frac{1}{2}\Delta_{\chi^2}^2(l_1, l_2),$$

where $\Delta_{\chi^2}^2(l_1, l_2) = \frac{(l_2-l_1)^2}{s^2}$. Thus it follows that $h_{f_{\chi}}(u) = \frac{1}{2}u^2$.

Now, since any f-divergence between any two Cauchy location densities $p_{l_1, s}$ and $p_{l_2, s}$ is a scalar function of the $\chi^2$-divergence [21]:

$$I_f(p_{l_1, s} : p_{l_2, s}) = g_f(\chi_s(l_1, l_2)),$$
we have
\[ I_f(p_{l_1,s} : p_{l_2,s}) = g_f \left( \frac{1}{2} \Delta^2_{s^2} \right). \]

Therefore it follows that
\[ h_f(u) = g_f \left( \frac{1}{2} u^2 \right). \]

See Table 1 of [21] for several examples of scalar functions corresponding to \( f \)-divergences.

Now we let \( d \geq 2 \). Contrary to the normal case, it is in general difficult to obtain explicit expressions for \( h_f \) in the Cauchy case. Here, we give one example to illustrate that difficulty. Let \( f(u) := (u - 1)^2 \) (the corresponding divergence is the \( \chi^2 \) divergence), \( d = 3 \), and \( F \) as Eq. (3). Then,
\[ F(t) = \int_{\mathbb{R}^3} p(||x - t||^2) dx - 1, \quad t \in \mathbb{R}^3. \]

By calculations, we find that
\[
\int_{\mathbb{R}} \frac{(1 + x_1^2 + x_2^2 + x_3^2)^2}{(1 + (x_1 - t)^2 + x_2^2 + x_3^2)^4} dx_1 = \frac{\pi}{16(1 + x_2^2 + x_3^2)^{3/2}} \left( 8 + \frac{16t^2}{1 + x_2^2 + x_3^2} + \frac{5t^4}{(1 + x_2^2 + x_3^2)^2} \right).
\]

Therefore,
\[
F(t) = \frac{1}{\pi} \int_{\mathbb{R}^2} \frac{\pi}{16(1 + x_2^2 + x_3^2)^{3/2}} \left( 8 + \frac{16t^2}{1 + x_2^2 + x_3^2} + \frac{5t^4}{(1 + x_2^2 + x_3^2)^2} \right) dx_2 dx_3 - 1
= \frac{2}{3} t^2 + \frac{t^4}{8}.
\]

Hence,
\[ h_f(s) = \frac{2}{3} s + \frac{s^2}{8}, \quad s \geq 0. \]

Contrary to the normal case (see Table 2), \( h_f \) for the Cauchy family is a polynomial. This holds also true for the case that \( d = 5, 7, \ldots \).

3.3 Multivariate \( f \)-divergences as equivalent univariate \( f \)-divergences

Ali and Silvey [11] further showed how to replace a \( d \)-dimensional \( f \)-divergence by an equivalent 1-dimensional \( f \)-divergence for the fixed covariance matrix normal distributions:

**Proposition 2** ([11], Section 6). Let \( \Delta_{\Sigma}(\mu_1, \mu_2) \) denote the Mahalanobis distance. Then we have
\[ I_f(p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) = I_f \left( p_{\mu_0, 1} : p_{\Delta_{\Sigma}(\mu_1, \mu_2), 1} \right). \]

We can show this assertion by the change of variable \( y = \frac{1}{\Delta_{\Sigma}(\mu_1, \mu_2)} (x - \mu_1)^\top \Sigma^{-1} (\mu_2 - \mu_1) \). Notice that \( \Delta_{\Sigma}(\mu_1, \mu_2) = \Delta_1(0, \Delta_{\Sigma}(\mu_1, \mu_2)) \), and therefore we can write:
\[
I_f(p_{\mu_1, \Sigma} : p_{\mu_2, \Sigma}) = h_f(\Delta_{\Sigma}(\mu_1, \mu_2)),
= h_f(\Delta_1(0, \Delta_{\Sigma}(\mu_1, \mu_2))),
= I_f \left( p_{\mu_0, 1} : p_{\Delta_{\Sigma}(\mu_1, \mu_2), 1} \right).
\]
Figure 1: Estimating the functions $h_f(\Delta^2)$ by Monte Carlo integration (y-axis is $\hat{h}_f(\Delta^2)$ for $s = 10^8$ MC samples) for the total variation distance (left) and the Jensen-Shannon divergence (right).

Property 2 yields a computationally efficient method to calculate stochastically the $f$-divergences when not known in closed form (e.g., the Jensen-Shannon divergence). We can estimate the $f$-divergence using $s$ samples $x_1, \ldots, x_s$ independently and identically distributed from a propositional distribution $r(x)$ as:

$$\hat{I}_f[p : q] = \frac{1}{s} \sum_{i=1}^{s} \frac{1}{r(x_i)} p(x_i) f\left(\frac{q(x_i)}{p(x_i)}\right).$$

Using we take the propositional distribution $r(x) = p(x)$. Estimating $f$-divergences between isotropic Gaussians requires $O(sd)$ time. Thus Proposition 2 allows to shave a factor $d$.

Notice that when $h_f$ is not available in closed-form, we can tabulate the function using Monte Carlo stochastic estimations of $\hat{h}_f(\Delta^2)$. Moreover, using symbolic regression software, we can fit a formula to the experimental tabulated data (see the plots in Figure 1): For example, for the JSD with $\Delta^2 \in [\frac{1}{2}, 5]$, we find that the function $\tilde{h}_{JS}(u) = \frac{2.06709u}{w+x} + 2.27508$ approximates well the underlying intractable function $h_{JS}(u)$ (relative mean error less than 0.1%). This techniques proves useful specially for bounded $f$-divergences like the total variation distance or the Jensen-Shannon divergence.

4 $f$-divergences in multivariate scale families

Consider the scale family $N_\mu = \{p_{\mu, \Sigma} : \Sigma \succ 0\}$ of $d$-variate MVNs with a prescribed location $\mu \in \mathbb{R}^d$. Ali and Silvey [1] proved that all $f$-divergences between any two multivariate normal distributions with prescribed mean is an increasing function of $|1 - \lambda_i|$’s, where the $\lambda_i$’s denote the eigenvalues of $\Sigma_2 \Sigma_1^{-1}$. This property can be proven by using the definition of $f$-divergences of Eq. 1.
and the fact that

\[
p_{\mu,\Sigma_2\Sigma_1^{-1}}(x) = \prod_{i=1}^{d} \frac{p_{\lambda_i,\Sigma_2\Sigma_1^{-1}}(x_i)}{p_{\mu,\Sigma_1}(x_i)}.
\]

Thus we have

\[
I_f(p_{\mu,\Sigma_1} : p_{\mu,\Sigma_2}) = I_f(p_{\mu,\Sigma_2} : p_{\mu,\Sigma_1}^{-1}) = E_f(|1 - \lambda_1|, \ldots, |1 - \lambda_d|),
\]

where \(E_f(\cdot)\) is a \(d\)-variate totally symmetric function (invariant to permutations of arguments). Therefore the \(f\)-divergences are spectral matrix divergences \([13]\). In particular, one interesting case is when \(E_f(\cdot)\) is a separable function:

\[
E_f(|1 - \lambda_1|, \ldots, |1 - \lambda_d|) = \sum_{i=1}^{d} e_{fKL}(|1 - \lambda_i|).
\]

We shall illustrate these results with the Kullback-Leibler divergence and more generally with the \(\alpha\)-divergences \([2]\) or \(\alpha\)-Bhattacharyya divergences \([19]\) below:

- The Kullback-Leibler divergence: The well-known formula of the KLD between two same-mean MVNs is

\[
D_{\text{KL}}(p_{\mu,\Sigma_1} : p_{\mu,\Sigma_2}) = \frac{1}{2} \left( \log \det(\Sigma_2 \Sigma_1^{-1}) + \text{tr}(\Sigma_2 \Sigma_1^{-1} - I) \right).
\]

This expression can be rewritten as

\[
D_{\text{KL}}(p_{\mu,\Sigma_1} : p_{\mu,\Sigma_2}) = \sum_{i=1}^{d} e'_{\text{KL}}(\lambda_i(\Sigma_2 \Sigma_1^{-1})) \text{ where } e'_{\text{KL}}(v) = \frac{1}{2} \left( \log v + v - 1 \right),
\]

since \(\det(\Sigma_2 \Sigma_1^{-1}) = \prod_{i=1}^{d} \lambda_i(\Sigma_2 \Sigma_1^{-1})\) and \(\text{tr}(\Sigma_2 \Sigma_1^{-1} - I) = \sum_{i=1}^{d} (\lambda_i(\Sigma_2 \Sigma_1^{-1}) - 1)\). By a change of variable \(v = 1 - u\), we get

\[
e_{fKL}(u) = \frac{1}{2} \left( \log(1 - u) - u \right), \quad (4)
\]

and \(E_{fKL}(u_1, \ldots, u_d) = \sum_{i=1}^{d} e_{fKL}(u_i)\) (separable case). We check that the scalar function \(e_{fKL}(u)\) is an increasing function of \(u\).

- More generally, let us consider the family of \(\alpha\)-divergences \([2]\):

\[
D_{\alpha}(p : q) = \begin{cases} 
\frac{4}{1-\alpha^2} \left( 1 - \rho_{\frac{1-\alpha}{2}}(p : q) \right), & \alpha \notin \{-1, 1\} \\
D_{\text{KL}}(p : q), & \alpha = -1 \\
D_{\text{KL}}(q : p), & \alpha = 1.
\end{cases}
\]

where

\[
\rho_{\beta}(p : q) = \int p^\beta(x)q^{1-\beta}(x)d\mu(x),
\]

13
is the skew Bhattacharyya coefficient (a similarity measure also called an affinity measure). The skew Bhattacharyya distance \([19]\) is 
\[D_{\text{Bhat}}(p : q) = -\log \rho_\beta(p : q).\]
We have \(D_\alpha(q : p) = D_{-\alpha}(p : q).\) We recover the squared Hellinger divergence when \(\alpha = 0\) and the Neyman \(\chi^2\)-divergence when \(\alpha = 3\) (and the Pearson \(\chi^2\)-divergence when \(\alpha = -3\)). The \(\alpha\)-divergences are \(f\)-divergences for the following family \(f_\alpha(u)\) of generators:

\[
f_\alpha(u) = \begin{cases} \frac{4}{1-\alpha^2} \left( u - u^{\frac{4+\alpha}{2}} \right), & \alpha \not\in \{-1, 1\} \\ -\log u, & \alpha = -1 \\ u \log u, & \alpha = 1. \end{cases}
\]

We have the following closed-form formula between two scale normal distributions \([23]\) (page 46):

\[
\rho_\beta(p_\mu, \Sigma_1 : p_\mu, \Sigma_2) = \frac{\det((1 - \beta)\Sigma_1 + \beta\Sigma_2)^{\frac{1-\beta}{2}}}{\det((1 - \beta)\Sigma_1 + \beta\Sigma_2)^{\frac{\beta}{2}}}. \tag{5}
\]

We can rewrite Eq. 5 as follows:

\[
\det((1 - \beta)\Sigma_1 + \beta\Sigma_2) = \det(I + \beta(\Sigma_2\Sigma_1^{-1} - I)) \det(\Sigma_1),
\]

so that

\[
\rho_\beta(p_\mu, \Sigma_1 : p_\mu, \Sigma_2) = \frac{\det(\Sigma_2\Sigma_1^{-1})^{\frac{\beta}{2}}}{\det(I + \beta(\Sigma_2\Sigma_1^{-1} - I))^{\frac{\beta}{2}}}.\]

Using the eigenvalues \(\lambda_i\)'s for \(i \in \{1, \ldots, d\}\) of \(\Sigma_2\Sigma_1^{-1}\), we have

\[
\rho_\beta(p_\mu, \Sigma_1 : p_\mu, \Sigma_2) = \prod_{i=1}^{d} \sqrt{\frac{\lambda_i^{\beta}}{1 + \beta(\lambda_i - 1)}}.
\]

Indeed, consider the characteristic polynomial

\[
p_{\Sigma_2\Sigma_1^{-1}}(x) = \det \left( xI - \Sigma_2\Sigma_1^{-1} \right) = \prod_{i=1}^{d} (x - \lambda_i(\Sigma_2\Sigma_1^{-1})).
\]

We have \(\det(\Sigma_2\Sigma_1^{-1}) = p_{\Sigma_2\Sigma_1^{-1}}(0) = (-1)^d \prod_{i=1}^{d} \lambda_i\) and

\[
\det(I + \beta(\Sigma_2\Sigma_1^{-1} - I)) = \beta^d \det \left( \left( \frac{1}{\beta} - 1 \right) \Sigma_2\Sigma_1^{-1} \right) = \beta^d p_{\Sigma_2\Sigma_1^{-1}} \left( \frac{1}{\beta} - 1 \right).
\]

Thus the \(\alpha\)-divergences or the Bhattacharyya divergences are increasing functions of the \(|1 - \lambda_i|\)'s as stated by Ali and Silvey \([1]\).

Notice that the bounds on the total variation distance between two multivariate Gaussian distributions with same mean has been investigated in \([8]\) but no closed-form formula is known.

Finally, we show that the \(f\)-divergences between two densities of a scale family are always spectral matrix divergences:
Proposition 3. For location-scale families \( \{ p_{\mu, \Sigma}(x) = (\det \Sigma)^{-1/2} p(\Sigma^{-1/2}(x - \mu)) \}_{\mu, \Sigma} \) where \( p \) is the standard density such that \( p(x) = \tilde{p}(|x|^2) \) for some \( \tilde{p} \), every \( f \)-divergence \( I_f(p_{\mu_1, \Sigma_1} : p_{\mu_2, \Sigma_2}) \) between scale family is a function of the eigenvalues of \( \Sigma_1 \Sigma_2^{-1} \).

Proof. We can assume that \( \mu = 0 \). By the change-of-variable \( x = \Sigma_1^{1/2} y \),

\[
I_f(p_{\mu_1, \Sigma_1} : p_{\mu_2, \Sigma_2}) = \int_{\mathbb{R}^d} f \left( \frac{(\det \Sigma_1)^{1/2}}{(\det \Sigma_2)^{1/2}} \frac{\tilde{p}(\Sigma_1^{1/2} \Sigma_2 - 1 \Sigma_1^{1/2} y)}{\tilde{p}(|y|^2)} \right) \, dy.
\]

Since \( \Sigma_1 \) and \( \Sigma_2 \) are both symmetric matrices, \( \Sigma_1^{1/2} \) and \( \Sigma_2^{-1} \) are both symmetric, and hence, \( \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} \) is also symmetric. Hence it is diagonalizable by an orthogonal matrix and there exist real eigenvalues \( \lambda_1, \ldots, \lambda_d \) and \( \det(\Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2}) = \lambda_1 \cdots \lambda_d \). Since \( \Sigma_2 \) is positive-definite, \( \Sigma_2^{-1} \) is also positive-definite. By this and the fact that \( \Sigma_1^{1/2} \) is symmetric, \( \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} \) is positive-definite, and hence, \( \lambda_1, \ldots, \lambda_d \) are all positive. Now we recall that the set of eigenvalues of \( \Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} \) is identical with the set of eigenvalues of \( \Sigma_1^{1/2} \Sigma_2^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} = \Sigma_1 \Sigma_2^{-1} \), which is a well-known result in linear algebra. Hence,

\[
\int_{\mathbb{R}^d} f \left( \frac{(\det \Sigma_1)^{1/2}}{(\det \Sigma_2)^{1/2}} \frac{\tilde{p}(\Sigma_1^{1/2} \Sigma_2^{-1} \Sigma_1^{1/2} y)}{\tilde{p}(|y|^2)} \right) \, dy = \int_{\mathbb{R}^d} f \left( (\lambda_1 \cdots \lambda_d)^{1/2} \frac{\tilde{p}(\lambda_1 y_1^2 + \cdots + \lambda_d y_d^2)}{\tilde{p}(|y|^2)} \right) \, dy_1 \cdots dy_d.
\]

References

[1] Syed Muntaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society: Series B (Methodological), 28(1):131–142, 1966. [1] [2] [7] [11] [12] [14]

[2] Shun-Ichi Amari. Information Geometry and Its Applications. Applied Mathematical Sciences. Springer Japan, Tokyo, Japan, 2016. [13]

[3] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, Joydeep Ghosh, and John Lafferty. Clustering with Bregman divergences. Journal of machine learning research, 6(10), 2005. [6]

[4] Miquel Calvo and Josep Maria Oller. An explicit solution of information geodesic equations for the multivariate normal model. Statistics & Risk Modeling, 9(1-2):119–138, 1991. [9]

[5] Thomas M Cover. Elements of information theory. John Wiley & Sons, Hoboken, NJ, USA, 1999. [5] [6]
[6] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. Magyar Tud. Akad. Mat. Kutato Int. Koezl., 8:85–108, 1964.

[7] Jason Davis and Inderjit Dhillon. Differential entropic clustering of multivariate gaussians. Advances in Neural Information Processing Systems, 19, 2006.

[8] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians with the same mean. arXiv preprint arXiv:1810.08693, 2018.

[9] Bent Fuglede and Flemming Topsoe. Jensen-Shannon divergence and Hilbert space embedding. In International Symposium on Information Theory (ISIT), page 31. IEEE, 2004.

[10] Wolfgang Globke and Raul Quiroga-Barranco. Information geometry and asymptotic geodesics on the space of normal distributions. Information Geometry, 4(1):131–153, 2021.

[11] Mohammadali Khosravifard, Dariush Fooladivanda, and T Aaron Gulliver. Confliction of the convexity and metric properties in f-divergences. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 90(9):1848–1853, 2007.

[12] Tonu Kollo. Advanced multivariate statistics with matrices. Springer, Berlin/Heidelberg, Germany, 2005.

[13] Brian Kulis, Mátyás A Sustik, and Inderjit S Dhillon. Low-Rank Kernel Learning with Bregman Matrix Divergences. Journal of Machine Learning Research, 10(2), 2009.

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

[15] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. Proceedings of the National Institute of Sciences (Calcutta), 2:49–55, 1936.

[16] Joseph V Michalowicz, Jonathan M Nichols, and Frank Bucholtz. Calculation of differential entropy for a mixed Gaussian distribution. Entropy, 10(3):200–206, 2008.

[17] Geert Molenberghs and Emmanuel Lesaffre. Non-linear integral equations to approximate bivariate densities with given marginals and dependence function. Statistica Sinica, pages 713–738, 1997.

[18] Frank Nielsen. On information projections between multivariate elliptical and location-scale families. arXiv preprint arXiv:2101.03839, 2021.

[19] Frank Nielsen and Sylvain Boltz. The Burbea-Rao and Bhattacharyya centroids. IEEE Transactions on Information Theory, 57(8):5455–5466, 2011.

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

[21] Frank Nielsen and Kazuki Okamura. On f-divergences between Cauchy distributions. arXiv preprint arXiv:2101.12459, 2021.
[22] Esa Ollila, David E Tyler, Visa Koivunen, and H Vincent Poor. Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on signal processing*, 60(11):5597–5625, 2012. 7

[23] Leandro Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, Boca Raton, Florida, 2018. 14

[24] Mohammad H Rohban, Prakash Ishwar, Burkay Orten, William Clement Karl, and Venkatesh Saligrama. An impossibility result for high dimensional supervised learning. In *2013 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2013. 9

[25] AGM Steerneman and F Van Perlo-Ten Kleij. Spherical distributions: Schoenberg (1938) revisited. *Expositiones Mathematicae*, 23(3):281–287, 2005. 7