The statistical Minkowski distances: Closed-form formula for Gaussian Mixture Models

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

Abstract
The traditional Minkowski distances are induced by the corresponding Minkowski norms in real-valued vector spaces. In this work, we propose novel statistical symmetric distances based on the Minkowski’s inequality for probability densities belonging to Lebesgue spaces. These statistical Minkowski distances admit closed-form formula for Gaussian mixture models when parameterized by integer exponents. This result extends to arbitrary mixtures of exponential families with natural parameter spaces being cones: This includes the binomial, the multinomial, the zero-centered Laplacian, the Gaussian and the Wishart mixtures, among others. We also derive a Minkowski’s diversity index of a normalized weighted set of probability distributions from Minkowski’s inequality.

Keywords: Minkowski $\ell_p$ metrics, $L_p$ spaces, Minkowski’s inequality, statistical mixtures, exponential families, multinomial theorem, statistical divergence, information radius, projective distance, scale-invariant distance, homogeneous distance.

1 Introduction and motivation

1.1 Statistical distances between mixtures

Gaussian Mixture Models (GMMs) are flexible statistical models often used in machine learning, signal processing and computer vision [41, 19] since they can arbitrarily closely approximate any smooth density. To measure the dissimilarity between probability distributions, one often relies on the principled information-theoretic Kullback-Leibler (KL) divergence [8], commonly called the relative entropy. However the lack of closed-form formula for the KL divergence between GMMs\(^1\) has motivated various KL lower and upper bounds [16, 15, 37, 38] for GMMs or approximation techniques [10], and further spurred the design of novel distances that admit closed-form formula between GMMs [28]. To give a few examples, let us cite the statistical squared Euclidean distance [19, 21], the Jensen-Rényi divergence [41] (for the quadratic Rényi entropy), the

\(^1\)When the GMMs share the same components, it is known that the KL divergence between them amount to an equivalent Bregman divergence [35] that is however computationally intractable because its corresponding Bregman generator is the differential negentropy that does not admit a closed-form expression in that case.
Cauchy-Schwarz (CS) divergence [18, 20], and a statistical distance based on discrete optimal transport [22, 38].

A distance \( D : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is a non-negative real-valued function \( D \) on the product space \( \mathcal{X} \times \mathcal{X} \) such that \( D(p, q) = D((p, q)) = 0 \) iff. \( p = q \). A distance \( D(p : q) \) between \( p \) and \( q \) may not be symmetric: This fact is emphasized by the ‘:’ delimiter notation: \( D(p : q) \neq D(q : p) \). For example, the KL divergence is an oriented distance: \( \text{KL}(p : q) \neq \text{KL}(q : p) \). Two usual symmetrizations of the KL divergence are the Jeffreys’ divergence and the Jensen-Shannon divergence [27]. Informally speaking, a divergence is a smooth distance that allows one to define an information-geometric structure [2]. In other words, a divergence is a smooth premetric distance [9].

Recently, the Cauchy-Schwarz divergence [18] has been generalized to Hölder divergences [39]. These Cauchy and Hölder distances \( D(p : q) \) are said to be projective because \( D(\lambda p : \lambda q) = D(p : q) \) for any \( \lambda, \lambda' > 0 \). An important family of projective divergences for robust statistical inference are the \( \gamma \)-divergences [13, 33]. Interestingly, those projective distances do not require to handle normalized probability densities but only need to consider positive densities instead (handy in applications). The Hölder projective divergences do not admit closed-form formula for GMMs, except for the very special case of the CS divergence. The underlying reason is that the conjugate exponents \( \frac{1}{\alpha} + \frac{1}{\beta} = 1 \) of Hölder divergences would need to be both integers. This constraint yields \( \alpha = \beta = 1 \), giving the special case of the CS divergence (the other integer exponent case is in the limit when \( \alpha = 0 \) and \( \beta = \infty \)).

1.2 Minkowski distances and Lebesgue spaces

The renown Minkowski distances are norm-induced metrics [9] measuring distances between \( d \)-dimensional vectors \( x, y \in \mathbb{R}^d \) defined for \( \alpha \geq 1 \) by:

\[
M_\alpha(x, y) := \|x - y\|_\alpha := \left( \sum_{i=1}^{d} |x_i - y_i|^{\alpha} \right)^{\frac{1}{\alpha}},
\]

where the Minkowski norms are given by \( \|x\|_\alpha := \left( \sum_{i=1}^{d} |x_i|^{\alpha} \right)^{\frac{1}{\alpha}} \). The Minkowski norms can be extended to countably infinite-dimensional \( \ell_\alpha \) spaces of sequences (see [1], p. 68).

Let \((\mathcal{X}, \mathcal{F})\) be a measurable space where \( \mathcal{F} \) denotes the \( \sigma \)-algebra of \( \mathcal{X} \), and let \( \mu \) be a probability measure (with \( \mu(\mathcal{X}) = 1 \)) with full support \( \text{supp}(\mu) = \mathcal{X} \) (where \( \text{supp}(\mu) := \text{cl}(\{F \in \mathcal{F} : \mu(F) > 0\}) \) and \( \text{cl} \) denotes the set closure). Let \( \mathcal{F} \) be the set of all real-valued measurable functions defined on \( \mathcal{X} \). We define the Lebesgue space \([1]\) \( L_\alpha(\mu) \) for \( \alpha \geq 1 \) as follows:

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

The Minkowski distance [25] of Eq. [1] can be generalized to probability densities belonging to Lebesgue \( L_\alpha(\mu) \) spaces, to get the statistical Minkowski distance for \( \alpha \geq 1 \):

\[
M_\alpha(p, q) := \left( \int_{\mathcal{X}} |p(x) - q(x)|^\alpha d\mu(x) \right)^{\frac{1}{\alpha}}.
\]

\(^2\) Also called a contrast function in [11].

\(^3\) A Riemannian distance is not smooth but a squared Riemannian distance is smooth.
When $\alpha = 1$, we recover twice the Total Variation (TV) metric:

$$\text{TV}(p,q) := \frac{1}{2} \int |p(x) - q(x)| \, d\mu(x) = \frac{1}{2} \| p - q \|_{L_1(\mu)} = \frac{1}{2} M_1(p,q).$$ (4)

Notice that the statistical Minkowski distance does not admit closed-form formula in general because of the absolute value. The total variation is related to the probability of error in Bayesian statistical hypothesis testing [29].

In this work, we design novel distances based on the Minkowski’s inequality (triangle inequality for $L_\alpha(\mu)$), which proves that $\| \cdot \|_{L_\alpha(\mu)}$ is a norm (i.e., the $L_\alpha$-norm), so that the statistical Minkowski’s distance between functions of a Lebesgue space can be written as $M_\alpha(p,q) = \| p - q \|_{L_\alpha(\mu)}$. The space $L_\alpha(\mu)$ is a Banach space (i.e., complete normed linear space).

1.3 Paper outline

The paper is organized as follows: Section 2 defines the new Minkowski distances by measuring in various ways the tightness of the Minkowski’s inequality applied to probability densities. Section 3 proves that all these statistical Minkowski distances admit closed-form formula for mixture of exponential families with conic natural parameter spaces for integer exponents. In particular, this includes the case of Gaussian mixture models. Section 4 lists a few examples of common exponential families with conic natural parameter spaces. In Section 5, we define Minkowski’s diversity indices for a normalized weighted set of probability distributions. Finally, section 6 concludes this work and hints at perspectives.

2 Distances from the Minkowski’s inequality

Let us state Minkowski’s inequality:

**Theorem 1** (Minkowski’s inequality). For $p(x), q(x) \in L_\alpha(\mu)$ with $\alpha \in [1, \infty)$, we have the following Minkowski’s inequality:

$$\left( \int |p(x) + q(x)|^\alpha \, d\mu(x) \right)^{\frac{1}{\alpha}} \leq \left( \int |p(x)|^\alpha \, d\mu(x) \right)^{\frac{1}{\alpha}} + \left( \int |q(x)|^\alpha \, d\mu(x) \right)^{\frac{1}{\alpha}},$$ (5)

with equality holding only when $q(x) = 0$ (almost everywhere, a.e.), or when $p(x) = \lambda q(x)$ a.e. for $\lambda > 0$ for $\alpha > 1$.

The usual proof of Minkowski’s inequality relies on Hölder’s inequality [40, 39]. Following [39], we define distances by measuring in several ways the tightness of the Minkowski’s inequality. When clear from context, we shall write $\| \cdot \|_\alpha$ for short instead of $\| \cdot \|_{L_\alpha(\mu)}$. Thus Minkowski’s inequality writes as:

$$\| p + q \|_\alpha \leq \| p \|_\alpha + \| q \|_\alpha.$$ (6)

Minkowski’s inequality proves that the $L_\alpha$-spaces are normed vector spaces.

Notice that when $p(x)$ and $q(x)$ are probability densities (i.e., $\int p(x) \, d\mu(x) = \int q(x) \, d\mu(x) = 1$), Minkowski’s inequality becomes an equality iff. $p(x) = q(x)$ almost everywhere, for $\alpha > 1$. Thus we can define the following novel Minkowski’s distances between probability densities satisfying the identity of indiscernibles:
Definition 2 (Minkowski difference distance). For probability densities \( p, q \in L^\alpha(\mu) \), we define the Minkowski difference \( D_\alpha(\cdot, \cdot) \) distance for \( \alpha \in (1, \infty) \) as:

\[
D_\alpha(p, q):= \|p\|_\alpha + \|q\|_\alpha - \|p + q\|_\alpha \geq 0. \tag{7}
\]

Definition 3 (Minkowski log-ratio distance). For probability densities \( p, q \in L^\alpha(\mu) \), we define the Minkowski log-ratio distance \( L_\alpha(\cdot, \cdot) \) for \( \alpha \in (1, \infty) \) as:

\[
L_\alpha(p, q):= -\log \frac{\|p + q\|_\alpha}{\|p\|_\alpha + \|q\|_\alpha} = \log \frac{\|p\|_\alpha + \|q\|_\alpha}{\|p + q\|_\alpha} \geq 0. \tag{8}
\]

By construction, all these Minkowski distances are symmetric distances: Namely, \( M_\alpha(p, q) = M_\alpha(q, p) \), \( D_\alpha(p, q) = D_\alpha(q, p) \) and \( L_\alpha(p, q) = L_\alpha(q, p) \).

Notice that \( L_\alpha(p, q) \) is scale-invariant \footnote{Like any distance based on the log ratio of triangle inequality gap induced by a homogeneous norm.}: \( L_\alpha(\lambda p, \lambda q) = L_\alpha(p, q) \) for any \( \lambda > 0 \). Scale-invariance is a useful property in many signal processing applications. For example, the scale-invariant Itakura-Saito divergence (a Bregman divergence) has been successfully used in Nonnegative Matrix Factorization \[12\] (NMF). Distance \( D_\alpha(p, q) \) is homogeneous since \( D_\alpha(\lambda p, \lambda q) = |\lambda|D_\alpha(p, q) \) for any \( \lambda \in \mathbb{R} \) (and so is distance \( M_\alpha(p, q) \)).

3 Closed-form formula for statistical mixtures of exponential families

In this section, we shall prove that \( D_\alpha \) and \( L_\alpha \) between statistical mixtures are in closed-form for all integer exponents (and \( M_\alpha \) for all even exponents) for mixtures of exponential families with conic natural parameter spaces.

Let us first define the positively weighed geometric integral \( I \) of a set \( \{p_1, \ldots, p_k\} \) of \( k \) probability densities of \( L^\alpha(\mu) \) as:

\[
I(p_1, \ldots, p_k; \alpha_1, \ldots, \alpha_k) := \int_X p_1(x)^{\alpha_1} \cdots p_k(x)^{\alpha_k} d\mu(x), \quad \alpha \in \mathbb{R}^k_+. \tag{9}
\]

An exponential family \[7, 31\] \( \mathcal{E}_{t,\mu} \) is a set \( \{p_\theta(x)\} \) of probability densities wrt. \( \mu \) which densities can be expressed proportionally canonically as:

\[
p_\theta(x) \propto \exp(t(x)^\top \theta), \tag{10}
\]

where \( t(x) \) is a \( D \)-dimensional vector of sufficient statistics \[7\]. The term \( t(x)^\top \theta \) can be written equivalently as \( \langle t(x), \theta \rangle \), where \( \langle \cdot, \cdot \rangle \) denotes the scalar product on \( \mathbb{R}^D \). Thus the normalized probability densities of \( \mathcal{E}_{t,\mu} \) can be written as:

\[
p_\theta(x) = \exp \left( t(x)^\top \theta - F(\theta) \right), \tag{11}
\]

where

\[
F(\theta) := \log \int_X \exp(t(x)^\top \theta) d\mu(x), \tag{12}
\]
is called the log-partition function (also called cumulant function \[7\] or log-normalizer \[31\]). The natural parameter space is:

\[ \Theta := \left\{ \theta \in \mathbb{R}^D : \int_{\mathcal{X}} \exp(t(x)^\top \theta) d\mu(x) < \infty \right\}. \tag{13} \]

Many common distributions (Gaussians, Poisson, Beta, etc.) belong to exponential families in disguise \[7, 31\].

**Lemma 4.** For probability densities \( p_{\theta_1}, \ldots, p_{\theta_k} \) belonging to the same exponential family \( \mathcal{E}_{t,\mu} \), we have:

\[ I(p_{\theta_1}, \ldots, p_{\theta_k}; \alpha_1, \ldots, \alpha_k) = \exp \left( F \left( \sum_{i=1}^{k} \alpha_i \theta_i \right) - \sum_{i=1}^{k} \alpha_i F(\theta_i) \right) < \infty, \tag{14} \]

provided that \( \sum_{i=1}^{k} \alpha_i \theta_i \in \Theta \).

**Proof.**

\[ I(p_{\theta_1}, \ldots, p_{\theta_k}; \alpha_1, \ldots, \alpha_k) = \int \prod_{i=1}^{k} \left( \exp \left( \left( t(x)^\top \theta_i - F(\theta_i) \right) \right) \right) ^{\alpha_i} d\mu(x), \]

\[ = \int \exp \left( t(x)^\top \left( \sum_{i} \alpha_i \theta_i \right) - \sum_{i} \alpha_i F(\theta_i) \right) + \sum_{i} F(\sum_{i} \alpha_i \theta_i) - F \left( \sum_{i} \alpha_i \theta_i \right) \right) d\mu(x), \]

\[ = \exp \left( F \left( \sum_{i} \alpha_i \theta_i \right) - \sum_{i} \alpha_i F(\theta_i) \right) \int \exp \left( t(x)^\top \left( \sum_{i} \alpha_i \theta_i \right) - F \left( \sum_{i} \alpha_i \theta_i \right) \right) d\mu(x), \]

since \( \int_{\mathcal{X}} \exp \left( t(x)^\top \left( \sum_{i} \alpha_i \theta_i \right) - F(\sum_{i} \alpha_i \theta_i) \right) d\mu(x) = \int_{\mathcal{X}} p_{\sum_{i} \alpha_i \theta_i}(x) d\mu(x) = 1 \), provided that \( \bar{\theta} := \sum_{i} \alpha_i \theta_i \in \Theta \) (and \( p_{\bar{\theta}} \in \mathcal{E}_{t,\mu} \)).

In particular, the condition \( \sum_{i} \alpha_i \theta_i \in \Theta \) always holds when the natural parameter space \( \Theta \) is a cone. In the remainder, we shall call those exponential families with natural parameter space being a cone, Conic Exponential Families (CEFs) for short. Note that when \( \sum_{i} \alpha_i \theta_i \not\in \Theta \), the integral \( I(p_{\theta_1}, \ldots, p_{\theta_k}; \alpha_1, \ldots, \alpha_k) \) diverges (that is, \( I(p_{\theta_1}, \ldots, p_{\theta_k}; \alpha_1, \ldots, \alpha_k) = \infty \)).

Observe that for a CEF density \( p_{\theta}(x) \), we have \( p_{\theta}(x)^{\alpha} \) in \( L_\alpha(\mu) \) for any \( \alpha \in [1, \infty) \).

**Corollary 5.** We have \( I(p_{\theta_1}, \ldots, p_{\theta_k}; \alpha_1, \ldots, \alpha_k) = \exp \left( F \left( \sum_{i} \alpha_i \theta_i \right) - \sum_{i} \alpha_i F(\theta_i) \right) < \infty \) for probability densities belonging to the same exponential family with natural parameter space \( \Theta \) being a cone.

We also note in passing that \( I(p_1, \ldots, p_k; \alpha_1, \ldots, \alpha_k) < \infty \) for \( \alpha \in \mathbb{R}^k \) for probability densities belonging to the same exponential family with natural parameter space being an affine space (e.g., Poisson or isotropic Gaussian families \[32\]).
Let us define:
\[
J_F(\theta_1, \ldots, \theta_k; \alpha_1, \ldots, \alpha_k) := \sum_i \alpha_i F(\theta_i) - F\left(\sum_i \alpha_i \theta_i\right),
\]  
(15)

This quantity is called the \textit{Jensen diversity} \cite{30} when \(\alpha \in \Delta_k\) (the \((k - 1)\)-dimensional standard simplex), or Bregman information\(^5\) in \cite{3}. Although the Jensen diversity is non-negative when \(\alpha \in \Delta_k\), this Jensen diversity of Eq. \((15)\) maybe negative when \(\alpha \in \mathbb{R}^k_+\). When \(\alpha \in \mathbb{R}^k_+\), we thus call the Jensen diversity the \textit{generalized Jensen diversity}. It follows that we have:
\[
I(p_{\theta_1}, \ldots, p_{\theta_k}; \alpha_1, \ldots, \alpha_k) = \exp\left(-J_F(\theta_1, \ldots, \theta_k; \alpha_1, \ldots, \alpha_k)\right)
\]  
(16)

The CEFs include the Gaussian family, the Wishart family, the Binomial/multinomial family, etc. \cite{7 31 28}.

Let us consider a finite positive mixture \(\tilde{m}(x) = \sum_{i=1}^k w_ip_i(x)\) of \(k\) probability densities, where the weight vector \(w \in \mathbb{R}^k_+\) are not necessarily normalized to one.

\textbf{Lemma 6.} For a finite positive mixture \(\tilde{m}(x)\) with components belonging to the same CEF, \(\|\tilde{m}\|_{L_\alpha(\mu)}\) is finite and in closed-form, for any integer \(\alpha \geq 2\).

\textit{Proof.} Consider the multinomial expansion \(\tilde{m}(x)^\alpha\) obtained by applying the multinomial theorem \cite{3}:
\[
\tilde{m}(x)^\alpha = \sum_{\sum_{i=1}^k \alpha_i = \alpha \atop \alpha_i \in \mathbb{N}} \left(\begin{array}{c} \alpha \\ \alpha_1, \ldots, \alpha_k \end{array}\right) \prod_{j=1}^k (w_j p_j(x))^\alpha_j,
\]  
(17)

where
\[
\left(\begin{array}{c} \alpha \\ \alpha_1, \ldots, \alpha_k \end{array}\right) := \frac{\alpha!}{\alpha_1! \times \ldots \times \alpha_k!},
\]  
(18)

is the \textit{multinomial coefficient} \cite{4}. It follows that:
\[
\int \tilde{m}(x)^\alpha d\mu(x) = \sum_{\sum_{i=1}^k \alpha_i = \alpha \atop \alpha_i \in \mathbb{N}} \left(\begin{array}{c} \alpha \\ \alpha_1, \ldots, \alpha_k \end{array}\right) \left(\prod_{j=1}^k w_j^{\alpha_j}\right) I(p_1, \ldots, p_k; \alpha_1, \ldots, \alpha_k).
\]  
(19)

Thus the term \(\|\tilde{m}\|_{L_\alpha(\mu)}\) amounts to a positively weighted sum of integrals of monomials that are positively weighted geometric means of mixture components. When \(p_i = p_{\theta_i}\), since \(I(p_{\theta_1}, \ldots, p_{\theta_k}; \alpha_1, \ldots, \alpha_k) < \infty\) using Eq. \[5\] we conclude that \(\tilde{m} \in L_\alpha(\mu)\) for \(\alpha \in \mathbb{N}\), and we get the formula:
\[
\|\tilde{m}\|_{L_\alpha(\mu)} = \left(\sum_{\sum_{i=1}^k \alpha_i = \alpha \atop \alpha_i \in \mathbb{N}} \left(\begin{array}{c} \alpha \\ \alpha_1, \ldots, \alpha_k \end{array}\right) \left(\prod_{j=1}^k w_j^{\alpha_j}\right) \exp\left(-J_F(\theta_1, \ldots, \theta_k; \alpha_1, \ldots, \alpha_k)\right)\right)^\frac{1}{\alpha},
\]  
(20)

for \(\alpha \in \mathbb{N}\). \hfill \Box

\(^5\)Because \(\sum_i \alpha_i B_F(\theta_i : \bar{\theta}) = J_F(\theta_1, \ldots, \theta_k; \alpha_1, \ldots, \alpha_k)\) for the barycenter \(\bar{\theta} = \sum_i \alpha_i \theta_i\), where \(B_F(\theta : \theta') = F'(\theta) - F'(\theta') - (\theta - \theta')^\top \nabla F(\theta')\) is a Bregman divergence.
A naive multinomial expansion of $\tilde{m}(x)^\alpha$ yields $k^\alpha$ terms that can then be simplified. Using the multinomial theorem, there are $\binom{k+\alpha-1}{\alpha}$ integral terms in the formula of $\int (\sum_{i=1}^{k} w_i p_i(x))^\alpha d\mu(x)$. This number corresponds to the number of sequences of $k$ disjoint subsets whose union is $\{1, \ldots, \alpha\}$ (also called the number of ordered partitions but beware that some sets may be empty).

The multinomial expansion can be calculated efficiently using a generalization of Pascal’s triangle, called Pascal’s simplex [26], thus avoiding to compute from scratch all the multinomial coefficients.

We have the following generalized Pascal’s recurrence formula for calculating the multinomial coefficients:

$$\binom{\alpha}{\alpha_1, \ldots, \alpha_k} = \sum_{i=1}^{k} \binom{\alpha-1}{\alpha_i-1, \ldots, \alpha_k}.$$

(21)

with the terminal cases $\binom{\alpha}{\alpha_1, \ldots, \alpha_k} = 0$ if there exists $\alpha_i < 0$. Also by convention, we set conveniently $\binom{\alpha}{\alpha_1, \ldots, \alpha_k} = 0$ if there exists $\alpha_i > \alpha$.

An efficient way to implement the multinomial expansion using nested iterative loops follows from this identity:

$$\left( \sum_{i=1}^{k} x_i \right)^\alpha = \sum_{\alpha_1=0}^{\alpha} \sum_{\alpha_2=0}^{\alpha} \ldots \sum_{\alpha_k=0}^{\alpha} \binom{\alpha}{\alpha_1, \alpha_2, \ldots, \alpha_k} \cdot x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k}.$$

(22)

We are now ready to show when the statistical Minkowski’s distances $M_\alpha, D_\alpha$ and $L_\alpha$ are in closed-form for mixtures of CEFs using Lemma 6.

Theorem 7 (Closed-form formula for Minkowski’s distances). For mixtures $m = \sum_{i=1}^{k} w_i p_{\theta_i}$ and $m' = \sum_{j=1}^{k'} w_j' p_{\theta_j}'$ of CEFs $E_{\mu,t}$, $D_\alpha$ and $L_\alpha$ admits closed-form formula for integers $\alpha \geq 2$, and $M_\alpha$ is in closed-form when $\alpha \geq 2$ is an even positive integer.

Proof. For $D_\alpha$ and $L_\alpha$, it is enough to show that $\|m\|_{L_\alpha(\mu)}, \|m'\|_{L_\alpha(\mu)}$ and $\|m + m'\|_{L_\alpha(\mu)}$ are all in closed-form. This follows from Lemma 6 by setting $\tilde{m}$ to be $m$, $m'$ and $m + m'$, respectively. The overall number of generalized Jensen diversity terms in the formula of $D_\alpha$ or $L_\alpha$ is $O\left(\binom{k+k'+\alpha-1}{\alpha}\right)$.

Now, consider distance $M_\alpha$. To get rid of the absolute value in $M_\alpha$ for even integers $\alpha$, we rewrite $M_\alpha$ as follows:

$$M_\alpha(m, m') = \|m - m'\|_{L_\alpha(\mu)} = \left( \int |m(x) - m'(x)|^\alpha d\mu(x) \right)^{\frac{1}{\alpha}},$$

$$\left( \int \left((m(x) - m'(x))^2\right)^\frac{2}{\alpha} d\mu(x) \right)^{\frac{1}{\alpha}}.$$

Let $\tilde{m}(x) = (m(x) - m'(x))^2$. We have:

$$\tilde{m}(x) = (m(x) - m'(x))^2,$$

(23)

$$= m(x)^2 + m'(x)^2 - 2m(x)m'(x),$$

(24)

$$= \left( \sum_{i=1}^{k} w_i p_{\theta_i}(x) \right)^2 + \left( \sum_{j=1}^{k'} w_j' p_{\theta_j}'(x) \right)^2 - 2 \sum_{i=1}^{k} \sum_{j=1}^{k'} w_i w_j' p_{\theta_i}(x)p_{\theta_j}'(x).$$

(25)
We have the density products \( p_{\theta,\theta'} := p_{\theta} p_{\theta'} = I(p_{\theta}, p_{\theta'}; 1, 1) \in L_2(\mu) \) (using Lemma \( \square \) for any \( \theta, \theta' \in \Theta \) and \( \alpha \geq 2 \). When \( \alpha = 2, \frac{\alpha}{2} = 1 \), and we easily reach a closed-form formula for \( M_2(m, m') \). Otherwise, let us expand all the terms in Eq. \( \ref{eq:25} \) and rewrite \( \tilde{m}(x) = \sum_{l=1}^{K} w''_l p_{\theta_l, \theta'_l} \). Now, a key difference is that \( w''_l \in \mathbb{R} \), and not necessarily positive. Nevertheless, since \( \frac{\alpha}{2} \in \mathbb{N} \), we can still use the multinomial theorem to expand \( \tilde{m}(x)^{\frac{\alpha}{2}} \), distribute the integral over all terms, and compute elementary integrals \( I(p_{\theta_l, \theta'_1}, \ldots, p_{\theta_k, \theta'_k}; \alpha'_1, \ldots, \alpha'_K) \) with \( \sum_{l=1}^{K} \alpha'_l = \frac{\alpha}{2} \) in closed-form. Thus \( M_\alpha \) is available in closed-form for mixtures of CEFs for all even positive integers \( \alpha \geq 2 \). The number of terms in the \( M_\alpha \) formula is \( O\left((\max(k^2,k/2)+\alpha-1)\right) \). \( \square \)

Note that there exists a generalization\(^6\) of the binomial theorem to real exponents \( \alpha \in \mathbb{R} \) called \textit{Newton’s generalized binomial theorem} using an infinite series of general binomial coefficients:

\[
(x_1 + x_2)^\alpha = \sum_{i=0}^{\infty} \binom{\alpha}{i} x_1^i x_2^{\alpha-i},
\]

with the generalized binomial coefficient defined by:

\[
\binom{\alpha}{i} := \frac{\alpha(\alpha - 1) \ldots (\alpha - i + 1)}{i!} = \frac{\Gamma(\alpha + 1)}{\Gamma(\alpha - i + 1)\Gamma(i + 1)},
\]

where \( \Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt \) is the Gamma function extending the factorial: \( \Gamma(n) = (n-1)! \). Equation \( \ref{eq:26} \) is only valid whenever the infinite series converge. That is, for \( |x_1| \geq |x_2| \). When extending to mixture densities (i.e., \( (w_1 p_1(x) + w_2 p_2(x))^\alpha \)) and taking the integral, we therefore need to split the integral into two integrals depending on whether \( w_1 p_1(x) \geq w_2 p_2(x) \), or not. Furthermore, we need to compute these integrals on truncated support domains: This becomes very tricky as the dimension of the support increase \( \cite{14} \).

\section{Some examples of conic exponential families}

Let us report a few conic exponential families with their respective canonical decompositions. The measure \( \mu \) is usually either the Lebesgue measure on the Euclidean space (i.e., \( d\mu(x) = dx \)), or the counting measure.

- **Bernoulli/multinomial families.** The Bernoulli density is \( p(x; \lambda) = \lambda^x (1 - \lambda)^{1-x} \) with \( \lambda \in (0, 1) = \Delta_1 \), for \( \mathcal{X} = \{0, 1\} \). The natural parameter is \( \theta = \log \frac{1}{1-\lambda} \) and the conic natural parameter space is \( \Theta = \mathbb{R} \). The log-partition function is \( F(\theta) = \log(1 + e^\theta) \). The sufficient statistics is \( t(x) = x \).

  The multinomial density generalizes the Bernoulli and the binomial densities. Here, we consider the categorical distribution also called “multinoulli” distribution. The multinoulli density is given by:

\[
p(x; \lambda_1, \ldots, \lambda_d) = \prod_{i=1}^{d} \lambda_i^{x_i},
\]

\(^6\)There also exists a generalization of the multinomial theorem to real exponents, however, this is much less known in the literature (see \url{http://fractional-calculus.com/multinomial_theorem.pdf}).
Informally speaking, a diversity index is a quantity that measures the variability of elements in a data set (i.e., the diversity of a population). For example, the (sample) variance of a (finite) multivariate Gaussian family.

0. Zero-centered Laplacian family. The density is \( p(x; \sigma) = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}} \) and the sufficient statistic is \( t(x) = |x| \). The natural parameter is \( \theta = -\frac{1}{\sigma} \) with the conic parameter space \( \Theta = (-\infty, 0) = \mathbb{R}_{--} \). The log-normalizer is \( F(\theta) = \log(\frac{2}{\pi}) \). See [3] for an application of Laplacian mixtures.

0. Multivariate Gaussian family. The probability density of a \( d \) variate Gaussian distribution is:

\[
p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left( -\frac{(x - \mu)^T \Sigma^{-1} (x - \mu)}{2} \right), \quad x \in \mathbb{R}^d
\]

where \( |\Sigma| \) denotes the determinant of the positive-definite matrix \( \Sigma \). The natural parameter consists in a vector part \( \theta_v \) and a matrix part \( \theta_M \): \( \theta = (\theta_v, \theta_M) = (\Sigma^{-1} \mu, \Sigma^{-1}) \). The conic natural parameter space is \( \Theta = \mathbb{R}^d \times S^d_{++} \), where \( S^d_{++} \) denotes the cone of positive definite matrices of dimension \( d \times d \). The sufficient statistics are \( (x, xx^T) \). The log-partition function is:

\[
F(\theta) = \frac{1}{2} \theta_v^T \Sigma^{-1} \theta_v - \frac{1}{2} \log |\theta_M| + \frac{d}{2} \log 2\pi.
\]

0. Wishart family. The probability density is

\[
p(X; n, S) = \frac{|X|^{\frac{n-d-1}{2}} e^{-\frac{1}{2} \text{tr}(S^{-1}X)}}{2^{\frac{d(n-d)}{2}} |S|^\frac{n}{2} \Gamma_d \left( \frac{n}{2} \right)}, \quad X \in S^d_{++}
\]

with \( S > 0 \) denoting the scale matrix and \( n > d - 1 \) denoting the number of degrees of freedom, where \( \Gamma_d \) is the multivariate Gamma function:

\[
\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{j=1}^d \Gamma \left( x + (1 - j)/2 \right).
\]

\( \text{tr}(X) \) denotes the trace of matrix \( X \). The natural parameter is composed of a scalar \( \theta_s \) and a matrix part \( \theta_M \): \( \theta = (\theta_s, \theta_M) = (\frac{n-d-1}{2}, S^{-1}) \). The conic natural parameter space is \( \Theta = \mathbb{R}_+ \times S^d_{++} \). The sufficient statistics are \( (\log |X|, X) \). The log-partition function is:

\[
F(\theta) = \left( \frac{2\theta_s + d + 1}{2} \right) \log 2 + \left( \theta_s + \frac{d + 1}{2} \right) \log |\theta_M| + \log \Gamma_d \left( \theta_s + \frac{d + 1}{2} \right).
\]

See [17] for an application of Wishart mixtures.

5 Minkowski’s diversity index

Informally speaking, a diversity index is a quantity that measures the variability of elements in a data set (i.e., the diversity of a population). For example, the (sample) variance of a (finite)
point set is a diversity index. Point sets uniformly filling a large volume have large variance (and a large diversity index) while point sets with points concentrating to their centers of mass have low variance (and a small diversity index).

Recall that the Jensen diversity index [34] of a normalized weighted set \( \{ p_1 = p_{\theta_1}, \ldots, p_n = p_{\theta_n} \} \) of densities belonging to the same exponential family (also called information radius [23] or Bregman information [5, 36]) is defined for a strictly convex generator \( F \) by:

\[
J_F(\theta_1, \ldots, \theta_n; w_1, \ldots, w_n) := \sum_{i=1}^{n} w_i F(\theta_i) - F \left( \sum_{i=1}^{n} w_i \theta_i \right) \geq 0.
\]

When \( F(\theta) = \frac{1}{2} \langle \theta, \theta \rangle \), we recover from \( J_F \) the variance.

We shall consider finite mixtures [24, 5] with linearly independent component densities. Using Minkowski’s inequality iteratively for \( f_1, \ldots, f_n \in L_\alpha(\mu) \), we get:

\[
\left( \int \left| \sum_{i=1}^{n} f_i(x) \right|^{\alpha} d\mu(x) \right)^{\frac{1}{\alpha}} \leq \sum_{i=1}^{n} \left( \int |f_i(x)|^{\alpha} d\mu(x) \right)^{\frac{1}{\alpha}}.
\]

When \( \alpha > 1 \), equality holds when the \( f_i \)'s are proportional (a.e. \( \mu \)). By setting \( f_i = w_i p_i \), we define the Minkowski’s diversity index:

**Definition 8** (Minkowski’s diversity index). Define the Minkowski diversity index of \( n \) weighted probability densities of \( L_\alpha(\mu) \) for \( \alpha > 1 \) by:

\[
J_M^\alpha(p_1, \ldots, p_n; w_1, \ldots, w_n) := \sum_{i=1}^{n} w_i \left( \int p_i(x)^{\alpha} d\mu(x) \right)^{\frac{1}{\alpha}} - \left( \int \left| \sum_{i=1}^{n} w_i p_i(x) \right|^{\alpha} d\mu(x) \right)^{\frac{1}{\alpha}},
\]

\[
= \sum_{i=1}^{n} w_i \| p_i \|_\alpha - \left\| \sum_{i=1}^{n} w_i p_i \right\|_\alpha \geq 0.
\]

It follows a closed-form formula for the Minkowski’s diversity index of a weighted set of distributions (i.e., a mixture) belonging to the same CEF:

**Corollary 9.** The Minkowski’s diversity index of \( n \) weighted probability distributions belonging to the same conic exponential family is finite and admits a closed-form formula for any integer \( \alpha \geq 2 \).

### 6 Conclusion and perspectives

Designing novel statistical distances which admit closed-form formula for Gaussian mixture models is important for a wide range of applications in machine learning, computer vision and signal processing [18]. In this paper, we proposed to use the Minkowski’s inequality to design novel statistical symmetric Minkowski distances by measuring the tightness of the inequality either as an arithmetic difference or as a log-ratio of the left-hand-side and right-hand-side of the inequality. We showed that these novel statistical Minkowski distances yield closed-form formula for mixtures of exponential families with conic natural parameter spaces whenever the integer exponent \( \alpha \geq 2 \). In particular, this result holds for Gaussian mixtures, Bernoulli mixtures, Wishart mixtures, etc. We termed those families as Conic Exponential Families (CEF). We also reported a closed-form
formula for the ordinary statistical Minkowski distance for even positive integer exponents. Finally, we defined the Minkowski’s diversity index of a weighted population of probability distributions (a mixture), and proved that this diversity index admits a closed-form formula when the distributions belong to the same CEF.

Let us conclude by listing the formula of the statistical Minkowski distances for $\alpha = 2$ for comparison with the Cauchy-Schwarz (CS) divergence:

$$M_2(m_1, m_2) := \|m_1 - m_2\|_2,$$
$$D_2(m_1, m_2) := \|m_1 + m_2\|_2 - (\|m_1\|_2 + \|m_2\|_2),$$
$$L_2(m_1, m_2) := -\log \frac{\|m_1 + m_2\|_2}{\|m_1\|_2 + \|m_2\|_2},$$
$$CS(m_1, m_2) := -\log \frac{\|m_1m_2\|_1}{\|m_1\|_2 \|m_2\|_2} = -\log \frac{\langle m_1, m_2 \rangle_2}{\|m_1\|_2 \|m_2\|_2},$$

where $\langle f, g \rangle_2 = \int f(x)g(x)d\mu(x)$ for $f, g \in L_2(\mu)$. Note that for $\alpha = 2$, $L_2(\mu)$ is a Hilbert space when equipped with this inner product. We get closed-form formula for these statistical Minkowski’s distances between mixtures $m_1$ and $m_2$ of CEFs, as well as for the Cauchy-Schwarz divergence. All those statistical distances can be computed in quadratic time in the number of mixture components.

Selecting a proper divergence from a priori first principles for a given application is a paramount but difficult task [9]. Often one is left by checking experimentally the performances of a few candidate divergences in order to select the a posteriori ‘best’ one. We hope that these newly proposed statistical Minkowski’s distances, $D_\alpha$ and scale-invariant $L_\alpha$, will prove experimentally useful in a number of applications ranging from computer vision to machine learning and signal processing.

Additional material is available from

https://franknielsen.github.io/MinkowskiStatDist/

Acknowledgments: The author would like to thank Gaëtan Hadjeres for his careful reading and feedback.

References

[1] C. Alabiso and I. Weiss. A Primer on Hilbert Space Theory: Linear Spaces, Topological Spaces, Metric Spaces, Normed Spaces, and Topological Groups. UNITEXT for Physics. Springer International Publishing, 2014.

[2] Shun-ichi Amari. Information geometry and its applications. Springer, 2016.

[3] Tahir Amin, Mehmet Zeytinoglu, and Ling Guan. Application of Laplacian mixture model to image and video retrieval. IEEE Transactions on Multimedia, 9(7):1416–1429, 2007.

[4] Venkataramanan K. Balakrishnan. Introductory discrete mathematics. Courier Corporation, 2012.
[5] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[6] D. W. Bolton. The multinomial theorem. *The Mathematical Gazette*, 52(382):336–342, 1968.

[7] Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. IMS, 1986.

[8] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[9] Michel Marie Deza and Elena Deza. Encyclopedia of distances. In *Encyclopedia of Distances*, pages 1–583. Springer, 2009.

[10] J.-L. Durrieu, J.-Ph. Thiran, and Finnian Kelly. Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4833–4836. IEEE, 2012.

[11] Shinto Eguchi. Second order efficiency of minimum contrast estimators in a curved exponential family. *The Annals of Statistics*, pages 793–803, 1983.

[12] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009.

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

[14] Alan Genz and Frank Bretz. *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media, 2009.

[15] Jacob Goldberger and Hagai Aronowitz. A distance measure between GMMs based on the unscented transform and its application to speaker recognition. In *European Conference on Speech Communication and Technology (INTERSPEECH)*, pages 1985–1988, 2005.

[16] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan. An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. page 487, 2003.

[17] Leonard R. Haff, Peter T. Kim, J.-Y. Koo, and Richards. Minimax estimation for mixtures of Wishart distributions. *The Annals of Statistics*, 39(6):3417–3440, 2011.

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

[19] Bing Jian and Baba C. Vemuri. Robust point set registration using Gaussian mixture models. *IEEE Transactions on pattern analysis and machine intelligence (PAMI)*, 33(8):1633–1645, 2011.

[20] Kittipat Kampa, Erion Hasanbelliu, and Jose C. Principe. Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians. In *International Joint Conference on Neural Networks (IJCNN)*, pages 2578–2585, 2011.
[21] Meizhu Liu, Baba C Vemuri, Shun-ichi Amari, and Frank Nielsen. Shape retrieval using hierarchical total Bregman soft clustering. *IEEE Transactions on pattern analysis and machine intelligence (PAMI)*, 34(12):2407–2419, 2012.

[22] Zhu Liu and Qian Huang. A new distance measure for probability distribution function of mixture type. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 616–619. IEEE, 2000.

[23] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[24] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley, 2004.

[25] Hermann Minkowski. *Geometrie der Zahlen*, volume 40. 1910.

[26] Sarah Mousley, Nathan Schley, and Amy Shoemaker. Planar rook algebra with colors and Pascal’s simplex. *arXiv preprint arXiv:1211.0663*, 2012.

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

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

[29] Frank Nielsen. Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means. *Pattern Recognition Letters*, 42:25–34, 2014.

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

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

[32] 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, 2014.

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

[34] Frank Nielsen and Richard Nock. Generalizing skew Jensen divergences and Bregman divergences with comparative convexity. *IEEE Signal Processing Letters*, 24(8):1123–1127, 2017.

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

[36] Frank Nielsen, Paolo Piro, and Michel Barlaud. Bregman vantage point trees for efficient nearest neighbor queries. In *ICME*, pages 878–881. IEEE, 2009.
[37] Frank Nielsen and Ke Sun. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442, 2016.

[38] Frank Nielsen and Ke Sun. On the chain rule optimal transport distance. *arXiv preprint arXiv:1812.08113*, 2018.

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

[40] Elmer Tolsted. An elementary derivation of the Cauchy, Hölder, and Minkowski inequalities from Young’s inequality. *Mathematics Magazine*, 37(1):2–12, 1964.

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