Entropy **2014**, *16*, 2131-2145; doi:10.3390/e16042131

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Information Geometry of Positive Measures and Positive-Definite Matrices: Decomposable Dually Flat Structure

Shun-ichi Amari

RIKEN Brain Science Institute, Hirosawa 2-1, Wako-shi, Saitama 351-0198, Japan;
E-Mail: amari@brain.riken.jp; Tel.: +81-48-467-9669; Fax: +81-48-467-9687

Received: 14 February 2014; in revised form: 9 April 2014 / Accepted: 10 April 2014 / Published: 14 April 2014

Abstract: Information geometry studies the dually flat structure of a manifold, highlighted by the generalized Pythagorean theorem. The present paper studies a class of Bregman divergences called the (ρ, τ)-divergence. A (ρ, τ)-divergence generates a dually flat structure in the manifold of positive measures, as well as in the manifold of positive-definite matrices. The class is composed of decomposable divergences, which are written as a sum of componentwise divergences. Conversely, a decomposable dually flat divergence is shown to be a (ρ, τ)-divergence. A (ρ, τ)-divergence is determined from two monotone scalar functions, ρ and τ. The class includes the KL-divergence, α-, β- and (α, β)-divergences as special cases. The transformation between an affine parameter and its dual is easily calculated in the case of a decomposable divergence. Therefore, such a divergence is useful for obtaining the center for a cluster of points, which will be applied to classification and information retrieval in vision. For the manifold of positive-definite matrices, in addition to the dually flatness and decomposability, we require the invariance under linear transformations, in particular under orthogonal transformations. This opens a way to define a new class of divergences, called the (ρ, τ)-structure in the manifold of positive-definite matrices.

Keywords: information geometry; dually flat structure; decomposable divergence; (ρ, τ)-structure

1. Introduction

Information geometry, originated from the invariant structure of a manifold of probability distributions, consists of a Riemannian metric and dually coupled affine connections with respect to the metric [1]. A manifold having a dually flat structure is particularly interesting and important. In such a manifold, there are two dually coupled affine coordinate systems and a canonical divergence, which is a Bregman divergence. The highlight is given by the generalized Pythagorean theorem and projection theorem. Information geometry is useful not only for statistical inference, but also for machine learning, pattern recognition, optimization and even for neural networks. It is also related to the statistical physics of Tsallis q-entropy [2–4].

The present paper studies a general and unique class of decomposable divergence functions in $\boldsymbol{R}_{+}^{n}$, the manifold of n-dimensional positive measures. This is the (ρ, τ)-divergences, introduced by Zhang [5,6], from the point of view of "representation duality". They are Bregman divergences generating a dually flat structure. The class includes the well-known Kullback-Leibler divergence, α-divergence, β-divergence and (α, β)-divergence [1,7–9] as special cases. The merit of a decomposable Bregman divergence is that the $\boldsymbol{\theta}$-$\boldsymbol{\eta}$ Legendre transformation is computationally tractable, where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are two affine coordinates systems coupled by the Legendre transformation. When one uses a dually flat divergence to define the center of a cluster of elements, the center is easily given by the arithmetic mean of the dual coordinates of the elements [10,11]. However, we need to calculate its primal coordinates. This is the $\boldsymbol{\theta}$-$\boldsymbol{\eta}$ transformation. Hence, our new type of divergences has an advantage of calculating $\boldsymbol{\theta}$-coordinates for clustering and related pattern matching problems. The most general class of dually flat divergences, not necessarily decomposable, is further given in $\boldsymbol{R}_{+}^{n}$. They are the $(\boldsymbol{\rho}, \boldsymbol{\tau})$ divergence.

Positive-definite (PD) matrices appear in many engineering problems, such as convex programming, diffusion tensor analysis and multivariate statistical analysis [12–20]. The manifold, PD$_n$, of $n \times n$ PD matrices form a cone, and its geometry is by itself an important subject of research. If we consider the submanifold consisting of only diagonal matrices, it is equivalent to the manifold of positive measures. Hence, PD matrices can be regarded as a generalization of positive measures. There are many studies on geometry and divergences of the manifold of positive-definite matrices. We introduce a general class of dually flat divergences, the (ρ, τ)-divergence. We analyze the cases when a (ρ, τ)-divergence is invariant under the general linear transformations, $Gl(n)$, and invariant under the orthogonal transformations, $O(n)$. They not only include many well-known divergences of PD matrices, but also give new important divergences.

The present paper is organized as follows. Section 2 is preliminary, giving a short introduction to a dually flat manifold and the Bregman divergence. It also defines the cluster center due to a divergence. Section 3 defines the (ρ, τ)-structure in $\boldsymbol{R}_{+}^{n}$. It gives dually flat decomposable affine coordinates and a related canonical divergence (Bregman divergence). Section 4 is devoted to the (ρ, τ)-structure of the manifold, PD$_n$, of PD matrices. We first study the class of divergences that are invariant under $O(n)$. We further study a decomposable divergence that is invariant under $Gl(n)$. It coincides with the invariant divergence derived from zero-mean Gaussian distributions with PD covariance matrices. They not only include various known divergences, but new remarkable ones. Section 5 discusses a general class of non-decomposable flat divergences and miscellaneous topics. Section 6 is the conclusions.

2. Preliminaries to Information Geometry of Divergence

2.1. Dually Flat Manifold

A manifold is said to have the dually flat Riemannian structure, when it has two affine coordinate systems $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^n)$ and $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$ (with respect to two flat affine connections) together with two convex functions, $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$, such that the two coordinates are connected by the Legendre transformations:

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}), \quad \boldsymbol{\theta} = \nabla\varphi(\boldsymbol{\eta}), \tag{1}$$

where ∇ is the gradient operator. The Riemannian metric is given by:

$$(g_{ij}(\boldsymbol{\theta})) = \nabla\nabla\psi(\boldsymbol{\theta}), \quad \left(g^{ij}(\boldsymbol{\eta})\right) = \nabla\nabla\varphi(\boldsymbol{\eta}) \tag{2}$$

in the respective coordinate systems. A curve that is linear in the $\boldsymbol{\theta}$-coordinates is called a $\boldsymbol{\theta}$-geodesic, and a curve linear in the $\boldsymbol{\eta}$-coordinates is called an $\boldsymbol{\eta}$-geodesic.

A dually flat manifold has a unique canonical divergence, which is the Bregman divergence defined by the convex functions,

$$D[P:Q] = \psi\left(\boldsymbol{\theta}_P\right) + \varphi\left(\boldsymbol{\eta}_Q\right) - \boldsymbol{\theta}_P \cdot \boldsymbol{\eta}_Q, \tag{3}$$

where $\boldsymbol{\theta}_P$ is the $\boldsymbol{\theta}$-coordinates of P, $\boldsymbol{\eta}_Q$ is the $\boldsymbol{\eta}$-coordinates of Q and $\boldsymbol{\theta}_P \cdot \boldsymbol{\eta}_Q = \sum_i \left(\theta_P^i\right)\left(\eta_{Qi}\right)$, where θ_P^i and η_{Qi} are components of $\boldsymbol{\theta}_p$ and $\boldsymbol{\eta}_Q$, respectively. The Pythagorean and projection theorems hold in a dually flat manifold:

Pythagorean Theorem Given three points, P, Q, R, when the $\boldsymbol{\eta}$-geodesic connecting P and Q is orthogonal to the $\boldsymbol{\theta}$-geodesic connecting Q and R with respect to the Riemannian metric,

$$D[P:Q] + D[Q:R] = D[P:R]. \tag{4}$$

Projection Theorem Given a smooth submanifold, S, let P_S be the minimizer of divergence from P to S,

$$P_S = \min_{Q \in S} D[P:Q]. \tag{5}$$

Then, P_S is the $\boldsymbol{\eta}$-geodesic projection of P to S, that is the $\boldsymbol{\eta}$-geodesic connecting P and P_S is orthogonal to S.

We have the dual of the above theorems where $\boldsymbol{\theta}$- and $\boldsymbol{\eta}$-geodesics are exchanged and $D[P:Q]$ is replaced by its dual $D[Q:P]$.

2.2. Decomposable Divergence

A divergence, $D[P:Q]$, is said to be decomposable, when it is written as a sum of component-wise divergences,

$$D[P:Q] = \sum_{i=1}^{n} d\left(\theta_P^i, \theta_Q^i\right), \tag{6}$$

where θ_P^i and θ_Q^i are the components of $\boldsymbol{\theta}_P$ and $\boldsymbol{\theta}_Q$ and $d\left(\theta_P^i, \theta_Q^i\right)$ is a scalar divergence function.

An f-divergence:

$$D_f[P:Q] = \sum p_i f\left(\frac{q_i}{p_i}\right) \tag{7}$$

is a typical example of decomposable divergence in the manifold of probability distributions, where $P = (\boldsymbol{p})$ and $Q = (\boldsymbol{q})$ are two probability vectors with $\sum p_i = \sum q_i = 1$. A convex function, $\psi(\boldsymbol{\theta})$, is said to be decomposable, when it is written as:

$$\psi(\boldsymbol{\theta}) = \sum_{i=1}^{n} \tilde{\psi}\left(\theta^i\right) \tag{8}$$

by using a scalar convex function, $\tilde{\psi}(\theta)$. The Bregman divergence derived from a decomposable convex function is decomposable.

When $\psi(\boldsymbol{\theta})$ is a decomposable convex function, its Legendre dual is also decomposable. The Legendre transformation is given componentwise as:

$$\eta_i = \tilde{\psi}'\left(\theta_i\right), \tag{9}$$

where $'$ is the differentiation of a function, so that it is computationally tractable. Its inverse transformation is also componentwise,

$$\theta_i = \tilde{\varphi}'\left(\eta_i\right), \tag{10}$$

where $\tilde{\varphi}$ is the Legendre dual of $\tilde{\psi}$.

2.3. Cluster Center

Consider a cluster of points $P_1, \cdots, P_m$ of which $\boldsymbol{\theta}$-coordinates are $\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_m$ and $\boldsymbol{\eta}$-coordinates are $\boldsymbol{\eta}_1, \cdots, \boldsymbol{\eta}_m$. The center, R, of the cluster with respect to the divergence, $D[P:Q]$, is defined by:

$$R = \arg\min_{Q} \sum_{i=1}^{m} D\left[Q : P_i\right]. \tag{11}$$

By differentiating $\sum D\left[Q : P_i\right]$ by $\boldsymbol{\theta}$ (the $\boldsymbol{\theta}$-coordinates of Q), we have:

$$\nabla\psi\left(\boldsymbol{\theta}_R\right) = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\eta}_i. \tag{12}$$

Hence, the cluster-center theorem due to Banerjee *et al.* [10] follows; see also [11]:

Cluster-Center Theorem The $\boldsymbol{\eta}$-coordinates $\boldsymbol{\eta}_R$ of the cluster center are given by the arithmetic average of the $\boldsymbol{\eta}$-coordinates of the points in the cluster:

$$\boldsymbol{\eta}_R = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\eta}_i. \tag{13}$$

When we need to obtain the $\boldsymbol{\theta}$-coordinates of the cluster center, it is given by the $\boldsymbol{\theta}$-$\boldsymbol{\eta}$ transformation from $\boldsymbol{\eta}_R$,

$$\boldsymbol{\theta}_R = \nabla\varphi\left(\boldsymbol{\eta}_R\right). \tag{14}$$

However, in many cases, the transformation is computationally heavy and intractable when the dimensions of a manifold is large. The transformation is easy in the case of a decomposable divergence. This is motivation for considering a general class of decomposable Bregman divergences.

3. (ρ, τ) Dually Flat Structure in $\boldsymbol{R}_+^n$

3.1. (ρ, τ)-Coordinates of $\boldsymbol{R}_+^n$

Let $\boldsymbol{R}_+^n$ be the manifold of positive measures over n elements $x_1, \cdots, x_n$. A measure (or a weight) of x_i is given by:

$$\xi_i = m\left(x_i\right) > 0 \tag{15}$$

and $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n)$ is a distribution of measures. When $\sum \xi_i = 1$ is satisfied, it is a probability measure. We write:

$$\boldsymbol{R}_n^+ = \{\boldsymbol{\xi}\,|\xi_i > 0\} \tag{16}$$

and $\boldsymbol{\xi}$ forms a coordinate system of $\boldsymbol{R}_+^n$.

Let $\rho(\xi)$ and $\tau(\xi)$ be two monotonically increasing differentiable functions. We call:

$$\theta = \rho(\xi), \quad \eta = \tau(\xi) \tag{17}$$

the ρ- and τ-representations of positive measure ξ. This is a generalization of the $\pm\alpha$ representations [1] and was introduced in [5] for a manifold of probability distributions. See also [6].

By using these functions, we construct new coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ of $\boldsymbol{R}_+^n$. They are given, for $\boldsymbol{\theta} = (\theta^i)$ and $\boldsymbol{\eta} = (\eta_i)$, by componentwise relations,

$$\theta^i = \rho\left(\xi_i\right), \quad \eta_i = \tau\left(\xi_i\right). \tag{18}$$

They are called the ρ- and τ-representations of $\boldsymbol{\xi} \in \boldsymbol{R}_+^n$, respectively. We search for convex functions, $\psi_{\rho,\tau}(\boldsymbol{\theta})$ and $\varphi_{\rho,\tau}(\boldsymbol{\eta})$, which are Legendre duals to each other, such that $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are two dually coupled affine coordinate systems.

3.2. Convex Functions

We introduce two scalar functions of θ and η by:

$$\tilde{\psi}_{\rho,\tau}(\theta) = \int_0^{\rho^{-1}(\theta)} \tau(\xi)\rho'(\xi)d\xi, \tag{19}$$

$$\tilde{\varphi}_{\rho,\tau}(\eta) = \int_0^{\tau^{-1}(\eta)} \rho(\xi)\tau'(\xi)d\xi. \tag{20}$$

Then, the first and second derivatives of $\tilde{\psi}_{\rho,\tau}$ are:

$$\tilde{\psi}'_{\rho,\tau}(\theta) = \tau(\xi), \tag{21}$$

$$\tilde{\psi}''_{\rho,\tau}(\theta) = \frac{\tau'(\xi)}{\rho'(\xi)}. \tag{22}$$

Since $\rho'(\xi) > 0$, $\tau'(\xi) > 0$, we see that $\tilde{\psi}_{\rho,\tau}(\theta)$ is a convex function. So is $\tilde{\varphi}_{\rho,\tau}(\eta)$. Moreover, they are Legendre duals, because:

$$\tilde{\psi}_{\rho,\tau}(\theta) + \tilde{\varphi}_{\rho,\tau}(\eta) - \theta\eta = \int_0^{\xi} \tau(\xi)\rho'(\xi)d\xi + \int_0^{\xi} \rho(\xi)\tau'(\xi)d\xi - \rho(\xi)\tau(\xi) \tag{23}$$

$$= 0. \tag{24}$$

We then define two decomposable convex functions of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ by:

$$\psi_{\rho,\tau}(\boldsymbol{\theta}) = \sum \tilde{\psi}_{\rho,\tau}\left(\theta^i\right), \tag{25}$$

$$\varphi_{\rho,\tau}(\boldsymbol{\eta}) = \sum \tilde{\varphi}_{\rho,\tau}\left(\eta_i\right). \tag{26}$$

They are Legendre duals to each other.

3.3. (ρ, τ)-Divergence

The (ρ, τ)-divergence between two points, $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \boldsymbol{R}_n^+$, is defined by:

$$\begin{aligned} D_{\rho,\tau}\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] &= \psi_{\rho,\tau}(\boldsymbol{\theta}) + \varphi_{\rho,\tau}(\boldsymbol{\eta}') - \boldsymbol{\theta} \cdot \boldsymbol{\eta}' && (27) \\ &= \sum_{i=1}^{n} \left[\int_0^{\xi_i} \tau(\xi)\rho'(\xi)d\xi + \int_0^{\xi_i'} \rho(\xi)\tau'(\xi)d\xi - \rho\left(\xi_i\right)\tau\left(\xi_i'\right)\right], && (28) \end{aligned}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\eta}'$ are ρ- and τ-representations of $\boldsymbol{\xi}$ and $\boldsymbol{\xi}'$, respectively.

The (ρ, τ)-divergence gives a dually flat structure having $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ as affine and dual affine coordinate systems. This is originally due to Zhang [5] and a generalization of our previous results concerning the q and deformed exponential families [4]. The transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is simple in the (ρ, τ)-structure, because it can be done componentwise,

$$\theta^i = \rho\left\{\tau^{-1}\left(\eta_i\right)\right\}, \tag{29}$$

$$\eta_i = \tau\left\{\rho^{-1}\left(\theta^i\right)\right\}. \tag{30}$$

The Riemannian metric is:

$$g_{ij}(\boldsymbol{\xi}) = \frac{\tau'\left(\xi_i\right)}{\rho'\left(\xi_i\right)}\delta_{ij}, \tag{31}$$

and hence Euclidean, because the Riemann-Christoffel curvature due to the Levi-Civita connection vanishes, too.

The following theorem is new, characterizing the (ρ, τ)-divergence.

Theorem 1. The (ρ, τ)-divergences form a unique class of divergences in $\boldsymbol{R}_+^n$ that are dually flat and decomposable.

3.4. Biduality: α-(ρ, τ) Divergence

We have dually flat connections, $\left(\nabla_{\rho,\tau}, \nabla^*_{\rho,\tau}\right)$, represented in terms of covariant derivatives, which are derived from $D_{\rho,\tau}$. This is called the representation duality by Zhang [5]. We further have the α-(ρ, τ) connections defined by:

$$\nabla^{(\alpha)}_{\rho,\tau} = \frac{1+\alpha}{2}\nabla_{\rho,\tau} + \frac{1-\alpha}{2}\nabla^*_{\rho,\tau}. \tag{32}$$

The α-$(-\alpha)$ duality is called the reference duality [5]. Therefore, $\nabla^{(\alpha)}_{\rho,\tau}$ possesses the biduality, one concerning α and $(-\alpha)$, and the other with respect to ρ and τ.

The Riemann-Christoffel curvature of $\nabla^{(\alpha)}_{\rho,\tau}$ is:

$$R^{(\alpha)}_{\rho,\tau} = \frac{1-\alpha^2}{4}R^{(0)}_{\rho,\tau} = 0 \tag{33}$$

for any α. Hence, there exists unique canonical divergence $D^{(\alpha)}_{\rho,\tau}$ and α-(ρ, τ) affine coordinate systems. It is an interesting future problem to obtain their explicit forms.

3.5. Various Examples

As a special case of the (ρ, τ)-divergence, we have the (α, β)-divergence obtained from the following power functions,

$$\rho(\xi) = \frac{1}{\alpha}\xi^{\alpha}, \; \tau(\xi) = \frac{1}{\beta}\xi^{\beta}. \tag{34}$$

This was introduced by Cichocki, Cruse and Amari in [7,8].

The affine and dual affine coordinates are:

$$\theta^i = \frac{1}{\alpha}\left(\xi_i\right)^{\alpha}, \quad \eta_i = \frac{1}{\beta}\left(\xi_i\right)^{\beta} \tag{35}$$

and the convex functions are:

$$\psi(\boldsymbol{\theta}) = c_{\alpha,\beta} \sum \theta_i^{\frac{\alpha+\beta}{\alpha}}, \quad \varphi(\boldsymbol{\eta}) = c_{\beta,\alpha} \sum \eta_i^{\frac{\alpha+\beta}{\beta}}, \tag{36}$$

where:

$$c_{\alpha,\beta} = \frac{1}{\beta(\alpha+\beta)}\alpha^{\frac{\alpha+\beta}{\alpha}}. \tag{37}$$

The induced (α, β)-divergence has a simple form,

$$D_{\alpha,\beta}[\boldsymbol{\xi} : \boldsymbol{\xi}'] = \frac{1}{\alpha\beta\left(\alpha+\beta\right)} \sum \left\{ \alpha\xi_i^{\alpha+\beta} + \beta\xi_i'^{\alpha+\beta} - (\alpha+\beta)\xi_i^{\alpha}\xi_i'^{\beta} \right\}, \tag{38}$$

for $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \boldsymbol{R}^n_+$. It is defined similarly in the manifold, S_n, of probability distributions, but it is not a Bregman divergence in S_n. This is because the total mass constraint $\sum \xi_i = 1$ is not linear in $\boldsymbol{\theta}$- or $\boldsymbol{\eta}$-coordinates in general.

The α-divergence is a special case of the (α, β)-divergence, so that it is a (ρ, τ)-divergence. By putting:

$$\rho(\xi) = \frac{2}{1-\alpha}\xi^{\frac{1-\alpha}{2}}, \quad \tau(\xi) = \frac{2}{1+\alpha}\xi^{\frac{1+\alpha}{2}}, \tag{39}$$

we have:

$$D_{\alpha}\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] = \frac{4}{1-\alpha^2} \sum \left\{ \frac{1-\alpha}{2}\xi_i + \frac{1+\alpha}{2}\xi_i^{\frac{1-\alpha}{2}} - \xi_i^{\alpha}\left(\xi_i'\right)^{\frac{1+\alpha}{2}} \right\}. \tag{40}$$

The β-divergence [19] is obtained from:

$$\rho(\xi) = \xi, \quad \tau(\xi) = \frac{1}{\beta}\xi^{1+\beta}. \tag{41}$$

It is written as:

$$D_{\beta}\left[\boldsymbol{\xi} : \boldsymbol{\xi}'\right] = \frac{1}{\beta(\beta+1)} \sum_i \left[\xi_i^{\beta+1} + (\beta+1)\xi_i' - \left(\xi_i'\right)^{\beta+1} - (\beta+1)\xi_i\left(\xi_i'\right)^{\beta} \right]. \tag{42}$$

The β-divergence is special in the sense that it gives a dually flat structure, even in S_n. This is because $u(\xi)$ is linear in ξ.

The classes of α-divergences and β-divergences intersect at the KL-divergence, and their duals are different in general. They are the only intersecting points of the two classes.

When $\rho(\xi) = \xi$ and $\tau(\xi) = U'(\xi)$ where U is a convex function, (ρ, τ)-divergence is Eguchi's U-divergence [21].

Zhang already introduced the (α, β)-divergence in [5], which is not a (ρ, τ)-divergence in $\boldsymbol{R}^n_+$ and different from ours. We regret for our confusing the naming of the (α, β)-divergence.

4. Invariant, Flat Decomposable Divergences in the Manifold of Positive-Definite Matrices

4.1. Invariant and Decomposable Convex Function

Let $\mathbf{P}$ be a positive-definite matrix and $\psi(\mathbf{P})$ be a convex function. Then, a Bregman divergence is defined between two positive definite matrices, $\mathbf{P}$ and $\mathbf{Q}$, by:

$$D[\mathbf{P} : \mathbf{Q}] = \psi(\mathbf{P}) - \psi(\mathbf{Q}) - \nabla\psi(\mathbf{P}) \cdot (\mathbf{P} - \mathbf{Q}) \tag{43}$$

where ∇ is the gradient operator with respect to matrix $\mathbf{P} = (P_{ij})$, so that $\nabla\psi(\mathbf{P})$ is a matrix and the inner product of two matrices is defined by:

$$\nabla\psi(\mathbf{Q}) \cdot \mathbf{P} = \mathrm{tr}\left\{\nabla\psi(\mathbf{Q})\mathbf{P}\right\}, \tag{44}$$

where tr is the trace of a matrix.

It induces a dually flat structure to the manifold of positive-definite matrices, where the affine coordinate system ($\boldsymbol{\theta}$-coordinates) is $\boldsymbol{\Theta} = \mathbf{P}$ and the dual affine coordinate system ($\boldsymbol{\eta}$-coordinates) is:

$$\mathbf{H} = \nabla\psi(\mathbf{P}). \tag{45}$$

A convex function, $\psi(\mathbf{P})$, is said to be invariant under the orthogonal group $O(n)$, when:

$$\psi(\mathbf{P}) = \psi\left(\mathbf{O}^T\mathbf{P}\mathbf{O}\right) \tag{46}$$

holds for any orthogonal transformation $\boldsymbol{O}$, where $\boldsymbol{O}^T$ is the transpose of $\boldsymbol{O}$. An invariant function is written as a symmetric function of n eigenvalues $\lambda_1, \cdots, \lambda_n$ of $\mathbf{P}$. See Dhillon and Tropp [12]. When an invariant convex function of $\mathbf{P}$ is written, by using a convex function, f, of one variable, in the additive form:

$$\psi(\mathbf{P}) = \sum f\left(\lambda_i\right), \tag{47}$$

it is said to be decomposable. We have:

$$\psi(\boldsymbol{P}) = \mathrm{tr} f(\mathbf{P}). \tag{48}$$

4.2. Invariant, Flat and Decomposable Divergence

A divergence $D[\mathbf{P} : \mathbf{Q}]$ is said to be invariant under $O(n)$, when it satisfies:

$$D[\mathbf{P} : \mathbf{Q}] = D\left[\mathbf{O}^T\mathbf{P}\mathbf{O} : \mathbf{O}^T\mathbf{Q}\mathbf{O}\right]. \tag{49}$$

When it is derived from a decomposable convex function, $\psi(\mathbf{P})$, it is invariant, flat and decomposable.

We give well-known examples of decomposable convex functions and the divergences derived from them:

(1) For $f(\lambda) = (1/2)\lambda^2$, we have:

$$\psi(\mathbf{P}) = \frac{1}{2}\sum \lambda_i^2, \tag{50}$$

$$D[\mathbf{P} : \mathbf{Q}] = \frac{1}{2}\|\mathbf{P} - \mathbf{Q}\|^2, \tag{51}$$

where $\|\mathbf{P}\|^2$ is the Frobenius norm:

$$\|\mathbf{P}\|^2 = \sum P_{ij}^2. \tag{52}$$

(2) For $f(\lambda) = -\log \lambda$

$$\psi(\mathbf{P}) = -\log\left(\det|\mathbf{P}|\right), \tag{53}$$

$$D[\mathbf{P} : \mathbf{Q}] = \operatorname{tr}\left(\mathbf{P}\mathbf{Q}^{-1}\right) - \log\left(\det\left|\mathbf{P}\mathbf{Q}^{-1}\right|\right) - n. \tag{54}$$

The affine coordinate system is $\mathbf{P}$, and the dual coordinate system is $\mathbf{P}^{-1}$. The derived geometry is the same as that of multivariate Gaussian probability distributions with mean zero and covariance matrix $\mathbf{P}$.

(3) For $f(\lambda) = \lambda \log \lambda - \lambda$,

$$\psi(\mathbf{P}) = \operatorname{tr}\left(\mathbf{P}\log\mathbf{P} - \mathbf{P}\right), \tag{55}$$

$$D[\mathbf{P} : \mathbf{Q}] = \operatorname{tr}\left(\mathbf{P}\log\mathbf{P} - \mathbf{P}\log\mathbf{Q} - \mathbf{P} + \mathbf{Q}\right). \tag{56}$$

This divergence is used in quantum information theory. The affine coordinate system is $\mathbf{P}$, and the dual affine coordinate system is $\log\mathbf{P}$; and, $\psi(\mathbf{P})$ is called the negative von Neuman entropy.

4.3. (ρ, τ)-*Structure in Positive Definite Matrices*

We extend the (ρ, τ)-structure in the previous section to the matrix case and introduce a general dually flat invariant decomposable divergence in the manifold of positive-definite matrices. Let:

$$\mathbf{\Theta} = \rho(\mathbf{P}), \quad \mathbf{H} = \tau(\mathbf{P}) \tag{57}$$

be ρ- and τ-representations of matrices. We use two functions, $\tilde{\psi}_{\rho,\tau}(\theta)$ and $\tilde{\varphi}_{\rho,\tau}(\eta)$, defined in Equations (19) and (20), for defining a pair of dually coupled invariant and decomposable convex functions,

$$\psi(\mathbf{\Theta}) = \operatorname{tr}\tilde{\psi}_{\rho,\tau}\left\{\mathbf{\Theta}\right\}, \tag{58}$$

$$\varphi(\mathbf{H}) = \operatorname{tr}\tilde{\varphi}_{\rho,\tau}\left\{\mathbf{H}\right\}. \tag{59}$$

They are not convex with respect to $\mathbf{P}$, but are convex with respect to $\mathbf{\Theta}$ and $\mathbf{H}$, respectively. The derived Bregman divergence is:

$$D[\mathbf{P} : \mathbf{Q}] = \psi\left\{\mathbf{\Theta}(\mathbf{P})\right\} + \varphi\left\{\mathbf{H}(\mathbf{Q})\right\} - \mathbf{\Theta}(\mathbf{P}) \cdot \mathbf{H}(\mathbf{Q}). \tag{60}$$

Theorem 2. The (ρ, τ)-divergences form a unique class of invariant, decomposable and dually flat divergences in the manifold of positive matrices.

The Euclidean, Gaussian and von Neuman divergences given in Equations (51), (54) and (56) are special examples of (ρ, τ)-divergences. They are given, respectively, by:

$$(1) \qquad \rho(\xi) = \tau(\xi) = \xi, \tag{61}$$

$$(2) \qquad \rho(\xi) = \xi, \quad \tau(\xi) = -\frac{1}{\xi}, \tag{62}$$

$$(3) \qquad \rho(\xi) = \xi, \quad \tau(\xi) = \log \xi. \tag{63}$$

When ρ and τ are power functions, we have the (α, β)-structure in the manifold of positive-definite matrices.

(4) $(\alpha$-$\beta)$-divergence.

By using the (α, β) power functions given by Equation (34), we have:

$$\psi(\mathbf{\Theta}) = \frac{\alpha}{\alpha+\beta} \mathrm{tr}\, \mathbf{\Theta}^{\frac{\alpha+\beta}{\alpha}} = \frac{\alpha}{\alpha+\beta} \mathrm{tr}\, \mathbf{P}^{\alpha+\beta}, \tag{64}$$

$$\varphi(\mathbf{H}) = \frac{\beta}{\alpha+\beta} \mathrm{tr}\, \mathbf{H}^{\frac{\alpha+\beta}{\beta}} = \frac{\beta}{\alpha+\beta} \mathrm{tr}\, \mathbf{P}^{\alpha+\beta} \tag{65}$$

so that the (α, β)-divergence of matrices is:

$$D[\mathbf{P} : \mathbf{Q}] = \mathrm{tr} \left\{ \frac{\alpha}{\alpha+\beta} \mathbf{P}^{\alpha+\beta} + \frac{\beta}{\alpha+\beta} \mathbf{Q}^{\alpha+\beta} - \mathbf{P}^{\alpha} \mathbf{Q}^{\beta} \right\}. \tag{66}$$

This is a Bregman divergence, where the affine coordinate system is $\mathbf{\Theta} = \mathbf{P}^{\alpha}$ and its dual is $\mathbf{H} = \mathbf{P}^{\beta}$.

(5) The α-divergence is derived as:

$$\mathbf{\Theta}(\boldsymbol{P}) = \frac{2}{1-\alpha} \boldsymbol{P}^{\frac{1-\alpha}{2}}, \tag{67}$$

$$\psi(\mathbf{\Theta}) = \frac{2}{1+\alpha} \boldsymbol{P}, \tag{68}$$

$$D_{\alpha}[\mathbf{P} : \mathbf{Q}] = \frac{4}{1-\alpha^2} \mathrm{tr} \left(-\mathbf{P}^{\frac{1-\alpha}{2}} \mathbf{Q}^{\frac{1+\alpha}{2}} + \frac{1-\alpha}{2} \mathbf{P} + \frac{1+\alpha}{2} \mathbf{Q} \right). \tag{69}$$

The affine coordinate system is $\frac{2}{1-\alpha} \mathbf{P}^{\frac{1-\alpha}{2}}$, and its dual is $\frac{2}{1+\alpha} \mathbf{P}^{\frac{1+\alpha}{2}}$.

(6) The β-divergence is derived from Equation (41) as:

$$D_{\beta}[\mathbf{P} : \mathbf{Q}] = \frac{1}{\beta(\beta+1)} \mathrm{tr} \left[\mathbf{P}^{\beta+1} + (\beta+1)\mathbf{Q} - \mathbf{Q}^{\beta+1} - (\beta+1)\mathbf{P}\mathbf{Q}^{\beta} \right]. \tag{70}$$

4.4. Invariance Under $Gl(n)$

We extend the concept of invariance under the orthogonal group to that under the general linear group, $Gl(n)$, that is the set of invertible matrices, $\boldsymbol{L}$, $\det |\boldsymbol{L}| \neq 0$. This is a stronger condition. A divergence is said to be invariant under $Gl(n)$, when:

$$D[\mathbf{P} : \mathbf{Q}] = D \left[\mathbf{L}^T \mathbf{P} \mathbf{L} : \mathbf{L}^T \mathbf{Q} \mathbf{L} \right] \tag{71}$$

holds for any $\mathbf{L} \in Gl(n)$.

We identify matrix $\mathbf{P}$ with the zero-mean Gaussian distribution:

$$p(\boldsymbol{x}, \mathbf{P}) = \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\mathbf{P}^{-1}\boldsymbol{x} - \frac{1}{2}\log\det|\mathbf{P}| - c\right\}, \tag{72}$$

where c is a constant. We know that an invariant divergence belongs to the class of f-divergences in the case of a manifold of probability distributions, where the invariance means the geometry does not change under a one-to-one mapping of $\boldsymbol{x}$ to $\boldsymbol{y}$. Moreover, the only invariant flat divergence is the KL-divergence [22]. These facts suggest the following conjecture.

Proposition. The invariant, flat and decomposable divergence under $Gl(n)$ is the KL-divergence given by:

$$D_{KL}[\boldsymbol{P} : \boldsymbol{Q}] = \operatorname{tr}\left(\boldsymbol{P}\boldsymbol{Q}^{-1}\right) - \log\left(\det\left|\boldsymbol{P}\boldsymbol{Q}^{-1}\right|\right) - n. \tag{73}$$

5. Non-Decomposable Divergence

We have focused on flat and decomposable divergences. There are many interesting non-decomposable divergences. We first discuss a general class of flat divergences in $\boldsymbol{R}_+^n$ and then touch upon interesting flat and non-flat divergences in the manifold of positive-definite matrices.

5.1. General Class of Flat Divergences in $\boldsymbol{R}_+^n$

We can describe a general class of flat divergence in $\boldsymbol{R}_+^n$, which are not necessarily decomposable. This is introduced in [23], which studies the conformal structure of general total Bregman divergences ([11,13]). When $\boldsymbol{R}_+^n$ is endowed with a dually flat structure, it has a $\boldsymbol{\theta}$-coordinate system given by:

$$\boldsymbol{\theta} = \boldsymbol{\rho}(\boldsymbol{\xi}) \tag{74}$$

which is not necessarily a componentwise function. Any pair of invertible $\boldsymbol{\theta} = \boldsymbol{\rho}(\boldsymbol{\xi})$ and convex function $\psi(\boldsymbol{\theta})$ defines a dually flat structure and, hence, a Bregman divergence in $\boldsymbol{R}_+^n$.

The dual coordinates $\boldsymbol{\eta} = \boldsymbol{\tau}(\boldsymbol{\xi})$ are given by:

$$\boldsymbol{\eta} = \nabla\psi(\boldsymbol{\theta}) \tag{75}$$

so that we have:

$$\boldsymbol{\eta} = \boldsymbol{\tau}(\boldsymbol{\xi}) = \nabla\psi\left\{\boldsymbol{\rho}(\boldsymbol{\xi})\right\}. \tag{76}$$

This implies that a pair $(\boldsymbol{\rho}, \boldsymbol{\tau})$ of coordinate systems can define dually coupled affine coordinates and, hence, a dually flat structure, when and only when $\boldsymbol{\eta} = \boldsymbol{\tau}\left\{\boldsymbol{\rho}^{-1}(\boldsymbol{\theta})\right\}$ is a gradient of a convex function.

This is different from the case of decomposable divergence, where any monotone pair of $\rho(\xi)$ and $\tau(\xi)$ gives a dually flat structure.

5.2. Non-Decomposable Flat Divergence in PD_n

Ohara and Eguchi [15,16] introduced the following function:

$$\psi_V(\mathbf{P}) = V\left(\det|\mathbf{P}|\right), \tag{77}$$

where $V(\xi)$ is a monotonically decreasing scalar function. ψ_V is convex when and only when:

$$1 + \frac{V''(\xi)\xi^2}{V'(\xi)} < \frac{1}{n}. \tag{78}$$

In such a case, we can introduce dually flat structure to PD_n, where $\mathbf{P}$ is an affine coordinate system with convex $\psi_V(\mathbf{P})$, and the dual affine coordinate system is:

$$\mathbf{H} = V'(\det\|P\|)\mathbf{P}^{-1}. \tag{79}$$

The derived divergence is:

$$D_V[\mathbf{P}:\mathbf{Q}] = V(\det|\mathbf{P}) - V(\det|\mathbf{Q})| \tag{80}$$

$$+ V'(\det|\mathbf{Q}|)\mathrm{tr}\left\{\mathbf{Q}^{-1}(\mathbf{Q}-\mathbf{P})\right\}. \tag{81}$$

When $V(\xi) = -\log\xi$, it reduces to the case of Equation (54), which is invariant under $Gl(n)$ and decomposable. However, the divergence $D_V[\mathbf{P}:\mathbf{Q}]$ is not decomposable. It is invariant under $O(n)$ and more strongly so under $SGl(n) \subset Gl(n)$, defined by $\det|\mathbf{L}| = \pm 1$.

5.3. Flat Structure Derived from q-Escort Distribution

A dually flat structure is introduced in the manifold of probability distributions [4] as:

$$\tilde{D}_\alpha[\boldsymbol{p}:\boldsymbol{q}] = \frac{1}{1-q}\frac{1}{H_q(\boldsymbol{p})}\left(1 - \sum p_i^{1-q}q_i^q\right), \tag{82}$$

where:

$$H_q(\boldsymbol{p}) = \sum p_i^q, \tag{83}$$

$$q = \frac{1+\alpha}{2}. \tag{84}$$

The dual affine coordinates are the q-escort distribution: [4]

$$\eta_i = \frac{1}{H_q(\boldsymbol{p})}p_i^q. \tag{85}$$

The divergence, $\tilde{D}_q$, is flat, but not decomposable.

We can generalize it to the case of PD_n,

$$\tilde{D}_q[\mathbf{P}:\mathbf{Q}] = \frac{1}{1-q}\frac{1}{\mathrm{tr}\,\mathbf{P}^q}\left\{(1-q)\,\mathrm{tr}\,(\mathbf{P}) + q\,\mathrm{tr}\,(\mathbf{Q}) - \mathrm{tr}\,\left(\mathbf{P}^{1-\mathrm{q}}\mathbf{Q}^{\mathrm{q}}\right)\right\}. \tag{86}$$

This is flat, but not decomposable.

5.4. γ-Divergence in PD_n

The γ-divergence is introduced by Fujisawa and Eguchi [24]. It gives a super-robust estimator. It is interesting to generalize it to PD_n,

$$D_\gamma[\mathbf{P} : \mathbf{Q}] = \frac{1}{\gamma(\gamma-1)} \left\{ \log \mathrm{tr}\, \mathbf{P}^\gamma - (\gamma-1) \log \mathrm{tr}\, \mathbf{Q}^{\gamma-1} - \gamma \log \mathrm{tr}\, \mathbf{P}\mathbf{Q}^{\gamma-1} \right\}. \tag{87}$$

This is not flat nor decomposable. This is a projective divergence in the sense that, for any $c, c' > 0$,

$$D_\gamma \left[c\mathbf{P} : c'\mathbf{Q}\right] = D_\gamma[\mathbf{P} : \mathbf{Q}]. \tag{88}$$

Therefore, it can be defined in the submanifold of tr $\mathbf{P} = 1$.

6. Concluding Remarks

We have shown that the (ρ, τ)-divergence introduced by Zhang [5] is a general dually flat decomposable structure of the manifold of positive measures. We then extended it to the manifold of positive-definite matrices, where the criterion of invariance under linear transformations (in particular, under orthogonal transformations) were added. The decomposability is useful from the computational point of view, because the $\boldsymbol{\theta}$-$\boldsymbol{\eta}$ transformation is tractable. This is the motivation for studying decomposable flat divergences.

When we treat the manifold of probability distributions, it is a submanifold of the manifold of positive measures, where the total sum of measures are restricted to one. This is a nonlinear constraint in the $\boldsymbol{\theta}$ or $\boldsymbol{\eta}$ coordinates, so that the manifold is not flat, but curved in general. Hence, our arguments hold in this case only when at least one of the ρ and τ functions are linear. The U-divergence [21] and β-divergence [19] are such cases. However, for clustering, we can take the average of the $\boldsymbol{\eta}$-coordinates of member probability distributions in the larger manifold of positive measures and then project it to the manifold of probability distributions. This is called the exterior average, and the projection is simply a normalization of the result. Therefore, the (ρ, τ)-structure is useful in the case of probability distributions. The same situation holds in the case of positive-definite matrices.

Quantum information theory deals with positive-definite Hermitian matrices of trace one [25,26]. We need to extend our discussions to the case of complex matrices. The trace one constraint is not linear with respect to $\boldsymbol{\theta}$- or $\boldsymbol{\eta}$-coordinates, as is the same in the case of probability distributions. Many interesting divergence functions have been introduced in the manifold of positive-definite Hermitian matrices. It is an interesting future problem to apply our theory to quantum information theory.

Conflicts of Interest

The author declares no conflicts of interest.

References

1. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society and Oxford University Press: Rhode Island, RI, USA, 2000.

2. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*; Springer: Berlin/Heidelberg, Germany, 2009.
3. Naudts, J. *Generalized Thermostatistics*; Springer: Berlin/Heidelberg, Germany, 2011.
4. Amari, S.; Ohara, A.; Matsuzoe, H. Geometry of deformed exponential families: Invariant, dually-flat and conformal geometries. *Physica A* **2012**, *391*, 4308–4319.
5. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.
6. Zhang, J. Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy* **2013**, *15*, 5384–5418.
7. Cichocki, A.; Amari, S. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
8. Cichocki, A.; Cruces, S.; Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
9. Minami, M.; Eguchi, S. Robust blind source separation by beta-divergence. *Neural Comput.* **2002** *14*, 1859–1886.
10. Banerjee, A.; Merugu, S.; Dhillon I.; Ghosh, J. Clustering with Bregman Divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
11. Liu, M.; Vemuri, B.C.; Amari, S.; Nielsen, F. Shape retrieval using hierarchical total Bregman soft clustering. *IEEE Trans. Pattern Anal. Mach. Learn.* **2012**, *24*, 3192–3212.
12. Dhillon, I.S.; Tropp, J.A. Matrix nearness problems with Bregman divergences. *SIAM J. Matrix Anal. Appl.* **2007**, *29*, 1120–1146.
13. Vemuri, B.C.; Liu, M.; Amari, S.; Nielsen, F. Total Bregman divergence and its applications to DTI analysis. *IEEE Trans. Med. Imaging* **2011**, *30*, 475–483.
14. Ohara, A.; Suda, N.; Amari, S. Dualistic differential geometry of positive definite matrices and its applications to related problems. *Linear Algebra Appl.* **1996** *247*, 31–53.
15. Ohara, A.; Eguchi, S. Group invariance of information geometry on q-Gaussian distributions induced by beta-divergence. *Entropy* **2013**, *15*, 4732–4747.
16. Ohara, A.; Eguchi, S. Geometry on positive definite matrices induced from V-potential functions. In *Geometric Science of Information*; Nielsen, F., Barbaresco, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 621–629.
17. Chebbi, Z.; Moakher, M. Means of Hermitian positive-definite matrices based on the log-determinant alpha-divergence function. *Linear Algebra Appl.* **2012**, *436*, 1872–1889.
18. Tsuda, K.; Ratsch, G.; Warmuth, M.K. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *J. Mach. Learn. Res.* **2005**, *6*, 995–1018.
19. Nock, R.; Magdalou, B.; Briys, E.; Nielsen, F. Mining matrix data with Bregman matrix divergences for portfolio selection. In *Matrix Information Geometry*; Nielsen, F., Bhatia, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; Chapter 15, pp. 373–402.
20. Nielsen, F., Bhatia, R., Eds. *Matrix Information Geometry*; Springer: Berlin/Heidelberg, Germany, 2013.
21. Eguchi, S. Information geometry and statistical pattern recognition. *Sugaku Expo.* **2006**, *19*, 197–216.

22. Amari, S. α-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931.
23. Nock, R.; Nielsen, F.; Amari, S. On conformal divergences and their population minimizers. *IEEE Trans. Inf. Theory* **2014**, submitted for publication.
24. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081.
25. Petz, P. Monotone metrics on matrix spaces. *Linear Algebra Appl.* **1996**, *244*, 81–96.
26. Hasegawa, H. α-divergence of the non-commutative information geometry. *Rep. Math. Phys.* **1993**, *33*, 87–93.