Entropy **2015**, *17*, 2988-3034; doi:10.3390/e17052988

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Review

Log-Determinant Divergences Revisited: Alpha-Beta and Gamma Log-Det Divergences

Andrzej Cichocki [1,2,*], Sergio Cruces [3,*] and Shun-ichi Amari [4]

[1] Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, 2-1 Hirosawa, Wako, 351-0198 Saitama, Japan

[2] Systems Research Institute, Intelligent Systems Laboratory, Newelska 6, 01-447 Warsaw, Poland

[3] Dpto de Teoría de la Señal y Comunicaciones, University of Seville, Camino de los Descubrimientos s/n, 41092 Seville, Spain

[4] Laboratory for Mathematical Neuroscience, RIKEN BSI, Wako, 351-0198 Saitama, Japan; E-Mail: amari@brain.riken.jp

* Authors to whom correspondence should be addressed; E-Mails: a.cichocki@riken.jp (A.C.); sergio@us.es (S.C.).

Academic Editor: Raúl Alcaraz Martínez

Received: 19 December 2014 / Accepted: 5 May 2015 / Published: 8 May 2015

Abstract: This work reviews and extends a family of log-determinant (log-det) divergences for symmetric positive definite (SPD) matrices and discusses their fundamental properties. We show how to use parameterized Alpha-Beta (AB) and Gamma log-det divergences to generate many well-known divergences; in particular, we consider the Stein's loss, the S-divergence, also called Jensen-Bregman LogDet (JBLD) divergence, Logdet Zero (Bhattacharyya) divergence, Affine Invariant Riemannian Metric (AIRM), and other divergences. Moreover, we establish links and correspondences between log-det divergences and visualise them on an alpha-beta plane for various sets of parameters. We use this unifying framework to interpret and extend existing similarity measures for semidefinite covariance matrices in finite-dimensional Reproducing Kernel Hilbert Spaces (RKHS). This paper also shows how the Alpha-Beta family of log-det divergences relates to the divergences of multivariate and multilinear normal distributions. Closed form formulas are derived for Gamma divergences of two multivariate Gaussian densities; the special cases of the Kullback-Leibler, Bhattacharyya, Rényi, and Cauchy-Schwartz divergences are discussed. Symmetrized versions of log-det divergences are also considered and briefly reviewed. Finally, a class of divergences is extended to multiway divergences for separable covariance (or precision) matrices.

Keywords: Similarity measures; generalized divergences for symmetric positive definite (covariance) matrices; Stein's loss; Burg's matrix divergence; Affine Invariant Riemannian Metric (AIRM); Riemannian metric; geodesic distance; Jensen-Bregman LogDet (JBLD); S-divergence; LogDet Zero divergence; Jeffrey's KL divergence; symmetrized KL Divergence Metric (KLDM); Alpha-Beta Log-Det divergences; Gamma divergences; Hilbert projective metric and their extensions

1. Introduction

Divergences or (dis)similarity measures between symmetric positive definite (SPD) matrices underpin many applications, including: Diffusion Tensor Imaging (DTI) segmentation, classification, clustering, pattern recognition, model selection, statistical inference, and data processing problems [1–3]. Furthermore, there is a close connection between divergence and the notions of entropy, information geometry, and statistical mean [2,4–7], while matrix divergences are closely related to the invariant geometrical properties of the manifold of probability distributions [4,8–10]. A wide class of parameterized divergences for positive measures are already well understood and a unification and generalization of their properties can be found in [11–13].

The class of SPD matrices, especially covariance matrices, play a key role in many areas of statistics, signal/image processing, DTI, pattern recognition, and biological and social sciences [14–16]. For example, medical data produced by diffusion tensor magnetic resonance imaging (DTI-MRI) represent the covariance in a Brownian motion model of water diffusion. The diffusion tensors can be represented as SPD matrices, which are used to track the diffusion of water molecules in the human brain, with applications such as the diagnosis of mental disorders [14]. In array processing, covariance matrices capture both the variance and correlation of multidimensional data; this data is often used to estimate (dis)similarity measures, i.e., divergences. This all has led to an increasing interest in divergences for SPD (covariance) matrices [1,5,6,14,17–20].

The main aim of this paper is to review and extend log-determinant (log-det) divergences and to establish a link between log-det divergences and standard divergences, especially the Alpha, Beta, and Gamma divergences. Several forms of the log-det divergence exist in the literature, including the log–determinant α divergence, Riemannian metric, Stein's loss, S-divergence, also called the Jensen-Bregman LogDet (JBLD) divergence, and the symmetrized Kullback-Leibler Density Metric (KLDM) or Jeffrey's KL divergence [5,6,14,17–20]. Despite their numerous applications, common theoretical properties and the relationships between these divergences have not been established. To this end, we propose and parameterize a wide class of log-det divergences that provide robust solutions and/or even improve the accuracy for a noisy data. We next review fundamental properties and provide relationships among the members of this class. The advantages of some selected log-det divergences are also discussed; in particular, we consider the efficiency, simplicity, and resilience to noise or outliers, in addition to simplicity of calculations [14]. The log-det divergences between two SPD matrices have also been shown to be robust to biases in composition, which can cause problems for other similarity measures.

The divergences discussed in this paper are flexible enough to facilitate the generation of several established divergences (for specific values of the tuning parameters). In addition, by adjusting the adaptive tuning parameters, we optimize the cost functions of learning algorithms and estimate desired model parameters in the presence of noise and outliers. In other words, the divergences discussed in this paper are robust with respect to outliers and noise if the tuning parameters, α, β, and γ, are chosen properly.

1.1. Preliminaries

We adopt the following notation: SPD matrices will be denoted by $\mathbf{P} \in \mathbb{R}^{n\times n}$ and $\mathbf{Q} \in \mathbb{R}^{n\times n}$, and have positive eigenvalues λ_i (sorted in descending order); by $\log(\mathbf{P}),\ \det(\mathbf{P}) = |\mathbf{P}|,\ \mathrm{tr}(\mathbf{P})$ we denote the logarithm, determinant, and trace of $\mathbf{P}$, respectively.

For any real parameter $\alpha \in \mathbb{R}$ and for a positive definite matrix $\mathbf{P}$, the matrix $\mathbf{P}^\alpha$ is defined using symmetric eigenvalue decomposition as follows:

$$\mathbf{P}^\alpha = (\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T)^\alpha = \mathbf{V}(\boldsymbol{\Lambda}^\alpha)\,\mathbf{V}^T, \tag{1}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{P}$, and $\mathbf{V} \in \mathbb{R}^{n\times n}$ is the orthogonal matrix of the corresponding eigenvectors. Similarly, we define

$$\log(\mathbf{P}^\alpha) = \log((\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T)^\alpha) = \mathbf{V}\log(\boldsymbol{\Lambda}^\alpha)\,\mathbf{V}^T, \tag{2}$$

where $\log(\boldsymbol{\Lambda})$ is a diagonal matrix of logarithms of the eigenvalues of $\mathbf{P}$. The basic operations for positive definite matrices are provided in Appendix A.

The dissimilarity between two SPD matrices is called a metric if the following conditions hold:

1. $D(\mathbf{P}\,||\,\mathbf{Q}) \geq 0$, where the equality holds if and only if $\mathbf{P} = \mathbf{Q}$ (nonnegativity and positive definiteness).
2. $D(\mathbf{P}\,||\,\mathbf{Q}) = D(\mathbf{Q}\,||\,\mathbf{P})$ (symmetry).
3. $D(\mathbf{P}\,||\,\mathbf{Z}) \leq D(\mathbf{P}\,||\,\mathbf{Q}) + D(\mathbf{Q}\,||\,\mathbf{Z})$ (subaddivity/triangle inequality).

Dissimilarities that only satisfy condition (1) are not metrics and are referred to as (asymmetric) divergences.

2. Basic Alpha-Beta Log-Determinant Divergence

For SPD matrices $\mathbf{P} \in \mathbb{R}^{n\times n}$ and $\mathbf{Q} \in \mathbb{R}^{n\times n}$, consider a new dissimilarity measure, namely, the AB log-det divergence, given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \;=\; \frac{1}{\alpha\beta}\log\det\frac{\alpha(\mathbf{P}\mathbf{Q}^{-1})^{\beta} + \beta(\mathbf{P}\mathbf{Q}^{-1})^{-\alpha}}{\alpha+\beta} \tag{3}$$

$$\text{for } \alpha \neq 0,\quad \beta \neq 0,\quad \alpha+\beta \neq 0,$$

where the values of the parameters α and β can be chosen so as to guarantee the non-negativity of the divergence and it vanishes to zero if and only if $\mathbf{P} = \mathbf{Q}$ (this issue is addressed later by Theorems 1 and 2). Observe that this is not a symmetric divergence with respect to $\mathbf{P}$ and $\mathbf{Q}$, except when $\alpha = \beta$. Note that using the identity $\log\det(\mathbf{P}) = \mathrm{tr}\log(\mathbf{P})$, the divergence in (3) can be expressed as

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\alpha\beta}\,\mathrm{tr}\left[\log\left(\frac{\alpha(\mathbf{P}\mathbf{Q}^{-1})^{\beta} + \beta(\mathbf{P}\mathbf{Q}^{-1})^{-\alpha}}{\alpha+\beta}\right)\right] \tag{4}$$

$$\text{for} \quad \alpha \neq 0, \quad \beta \neq 0, \quad \alpha+\beta \neq 0.$$

This divergence is related to the Alpha, Beta, and AB divergences discussed in our previous work, especially Gamma divergences [11–13,21]. Furthermore, the divergence in (4) is related to the AB divergence for SPD matrices [1,12], which is defined by

$$\bar{D}_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\alpha\beta}\,\mathrm{tr}\left(\frac{\alpha}{\alpha+\beta}\mathbf{P}^{\alpha+\beta} + \frac{\beta}{\alpha+\beta}\,\mathbf{Q}^{\alpha+\beta} - \mathbf{P}^{\alpha}\mathbf{Q}^{\beta}\right) \tag{5}$$
$$\text{for} \quad \alpha \neq 0, \quad \beta \neq 0. \quad \alpha+\beta \neq 0.$$

Note that α and β are chosen so that $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ is nonnegative and equal to zero if $\mathbf{P} = \mathbf{Q}$. Moreover, such divergence functions can be evaluated without computing the inverses of the SPD matrices; instead, they can be evaluated easily by computing (positive) eigenvalues of the matrix $\mathbf{P}\mathbf{Q}^{-1}$ or its inverse. Since both matrices $\mathbf{P}$ and $\mathbf{Q}$ (and their inverses) are SPD matrices, their eigenvalues are positive. In general, it can be shown that even though $\mathbf{P}\mathbf{Q}^{-1}$ is nonsymmetric, its eigenvalues are the same as those of the SPD matrix $\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}$; hence, its eigenvalues are always positive.

Next, consider the eigenvalue decomposition:

$$(\mathbf{P}\mathbf{Q}^{-1})^{\beta} = \mathbf{V}\boldsymbol{\Lambda}^{\beta}\,\mathbf{V}^{-1}, \tag{6}$$

where $\mathbf{V}$ is a nonsingular matrix, and $\boldsymbol{\Lambda}^{\beta} = \mathrm{diag}\{\lambda_1^{\beta}, \lambda_2^{\beta}, \ldots, \lambda_n^{\beta}\}$ is the diagonal matrix with the positive eigenvalues $\lambda_i > 0,\ i = 1, 2, \ldots, n$, of $\mathbf{P}\mathbf{Q}^{-1}$. Then, we can write

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) &= \frac{1}{\alpha\beta}\log\det\frac{\alpha\,\mathbf{V}\boldsymbol{\Lambda}^{\beta}\,\mathbf{V}^{-1} + \beta\,\mathbf{V}\boldsymbol{\Lambda}^{-\alpha}\,\mathbf{V}^{-1}}{\alpha+\beta} \\ &= \frac{1}{\alpha\beta}\log\left[\det\mathbf{V}\ \det\frac{\alpha\boldsymbol{\Lambda}^{\beta} + \beta\boldsymbol{\Lambda}^{-\alpha}}{\alpha+\beta}\ \det\mathbf{V}^{-1}\right] \\ &= \frac{1}{\alpha\beta}\log\det\frac{\alpha\,\boldsymbol{\Lambda}^{\beta} + \beta\,\boldsymbol{\Lambda}^{-\alpha}}{\alpha+\beta}, \end{aligned} \tag{7}$$

which allows us to use simple algebraic manipulations to obtain

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) &= \frac{1}{\alpha\beta}\log\prod_{i=1}^{n}\frac{\alpha\lambda_i^{\beta} + \beta\lambda_i^{-\alpha}}{\alpha+\beta} \\ &= \frac{1}{\alpha\beta}\sum_{i=1}^{n}\log\left(\frac{\alpha\lambda_i^{\beta} + \beta\lambda_i^{-\alpha}}{\alpha+\beta}\right), \quad \alpha,\ \beta,\ \alpha+\beta \neq 0. \end{aligned} \tag{8}$$

It is straightforward to verify that $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = 0$ if $\mathbf{P} = \mathbf{Q}$. We will show later that this function is nonnegative for any SPD matrices if the α and β parameters take both positive or negative values.

For the singular values $\alpha = 0$ and/or $\beta = 0$ (also $\alpha = -\beta$), the AB log-det divergence in (3) is defined as a limit for $\alpha \to 0$ and/or $\beta \to 0$. In other words, to avoid indeterminacy or singularity for specific parameter values, the AB log-det divergence can be reformulated or extended by continuity and by applying L'Hôpital's formula to cover the singular values of α and β. Using L'Hôpital's rule, the AB log-det divergence can be defined explicitly by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \dfrac{1}{\alpha\beta}\log\det\dfrac{\alpha(\mathbf{P}\mathbf{Q}^{-1})^{\beta}+\beta(\mathbf{Q}\mathbf{P}^{-1})^{\alpha}}{\alpha+\beta} & \text{for } \alpha,\beta\neq 0,\ \ \alpha+\beta\neq 0 \\ \dfrac{1}{\alpha^2}\left[\operatorname{tr}\left((\mathbf{Q}\mathbf{P}^{-1})^{\alpha}-\mathbf{I}\right)-\alpha\log\det(\mathbf{Q}\mathbf{P}^{-1})\right] & \text{for } \ \alpha\neq 0,\ \beta=0 \\ \dfrac{1}{\beta^2}\left[\operatorname{tr}\left((\mathbf{P}\mathbf{Q}^{-1})^{\beta}-\mathbf{I}\right)-\beta\log\det(\mathbf{P}\mathbf{Q}^{-1})\right] & \text{for } \ \alpha=0,\ \beta\neq 0 \\ \dfrac{1}{\alpha^2}\log\dfrac{\det(\mathbf{P}\mathbf{Q}^{-1})^{\alpha}}{\det(\mathbf{I}+\log(\mathbf{P}\mathbf{Q}^{-1})^{\alpha})} & \text{for } \ \alpha=-\beta\neq 0 \\ \dfrac{1}{2}\operatorname{tr}\log^2(\mathbf{P}\mathbf{Q}^{-1}) = \dfrac{1}{2}||\log(\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})||_F^2 & \text{for } \ \alpha,\ \beta=0. \end{cases} \tag{9}$$

Equivalently, using standard matrix manipulations, the above formula can be expressed in terms of the eigenvalues of $\mathbf{P}\mathbf{Q}^{-1}$, *i.e.*, the generalized eigenvalues computed from $\lambda_i\mathbf{Q}\boldsymbol{v}_i = \mathbf{P}\boldsymbol{v}_i$ (where $\boldsymbol{v}_i$ ($i = 1,2,\ldots,n$) are corresponding generalized eigenvectors) as follows:

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \dfrac{1}{\alpha\beta}\displaystyle\sum_{i=1}^{n}\log\left(\dfrac{\alpha\lambda_i^{\beta}+\beta\lambda_i^{-\alpha}}{\alpha+\beta}\right) & \text{for } \alpha,\beta\neq 0, \quad \alpha+\beta\neq 0 \\ \dfrac{1}{\alpha^2}\left[\displaystyle\sum_{i=1}^{n}\left(\lambda_i^{-\alpha}-\log(\lambda_i^{-\alpha})\right)-n\right] & \text{for } \ \alpha\neq 0,\ \beta=0 \\ \dfrac{1}{\beta^2}\left[\displaystyle\sum_{i=1}^{n}\left(\lambda_i^{\beta}-\log(\lambda_i^{\beta})\right)-n\right] & \text{for } \ \alpha=0,\ \beta\neq 0 \\ \dfrac{1}{\alpha^2}\left[\displaystyle\sum_{i=1}^{n}\log\left(\dfrac{\lambda_i^{\alpha}}{1+\log\lambda_i^{\alpha}}\right)\right] & \text{for } \ \alpha=-\beta\neq 0 \\ \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\log^2(\lambda_i) & \text{for } \ \alpha,\ \beta=0. \end{cases} \tag{10}$$

Theorem 1. *The function $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \geq 0$ given by (3) is nonnegative for any SPD matrices with arbitrary positive eigenvalues if $\alpha \geq 0$ and $\beta \geq 0$ or if $\alpha < 0$ and $\beta < 0$. It is equal to zero if and only if $\mathbf{P} = \mathbf{Q}$.*

Equivalently, if the values of α and β have the same sign, the AB log-det divergence is positive independent of the distribution of the eigenvalues of $\mathbf{P}\mathbf{Q}^{-1}$ and goes to zero if and only if all the eigenvalues are equal to one. However, if the eigenvalues are sufficiently close to one, the AB log-det divergence is also positive for different signs of α and β. The conditions for positive definiteness are given by the following theorem.

Theorem 2. *The function* $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ *given by* (9) *is nonnegative if* $\alpha > 0$ *and* $\beta < 0$ *or if* $\alpha < 0$ *and* $\beta > 0$ *and if all the eigenvalues of* $\mathbf{PQ}^{-1}$ *satisfy the following conditions:*

$$\lambda_i \;>\; \left|\frac{\beta}{\alpha}\right|^{\frac{1}{\alpha+\beta}} \qquad \forall i, \textit{ for } \alpha > 0 \textit{ and } \beta < 0, \tag{11}$$

and

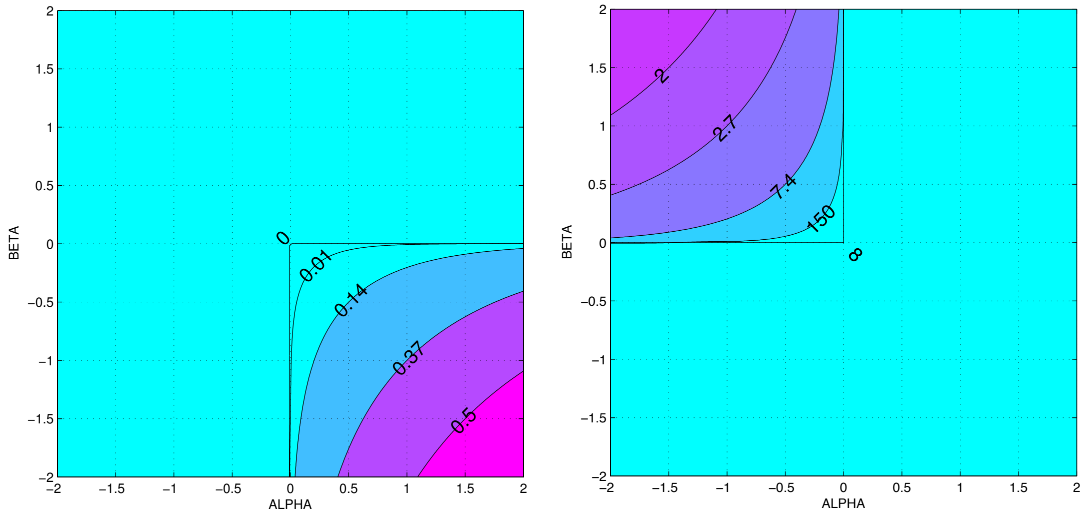
$$\lambda_i \;<\; \left|\frac{\beta}{\alpha}\right|^{\frac{1}{\alpha+\beta}} \qquad \forall i, \textit{ for } \alpha < 0 \textit{ and } \beta > 0. \tag{12}$$

If any of the eigenvalues violate these bounds, the value of the divergence, by definition, is infinite. Moreover, when $\alpha \to -\beta$ *these bounds simplify to*

$$\lambda_i \;>\; e^{-1/\alpha} \quad \forall i,\; \alpha = -\beta > 0, \tag{13}$$

$$\lambda_i \;<\; e^{-1/\alpha} \quad \forall i,\; \alpha = -\beta < 0. \tag{14}$$

In the limit, when $\alpha \to 0$ *or* $\beta \to 0$*, the bounds disappear. A visual presentation of these bounds for different values of* α *and* β *is shown in Figure 1.*
Additionally, $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = 0$ *only if* $\lambda_i = 1$ *for all* $i = 1, \ldots, n$*, i.e., when* $\mathbf{P} = \mathbf{Q}$.



(a) Lower-bounds of λ_i. (b) Upper-bounds of λ_i.

Figure 1. Shaded-contour plots of the bounds of λ_i that prevent $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ from diverging to ∞. The positive lower-bounds are shown in the lower-right quadrant of (**a**). The finite upper-bounds are shown in the upper-left quadrant of (**b**).

The proofs of these theorems are given in Appendices B, C and D.

Figure 2 illustrates the typical shapes of the AB log-det divergence for different values of the eigenvalues for various choices of α and β.

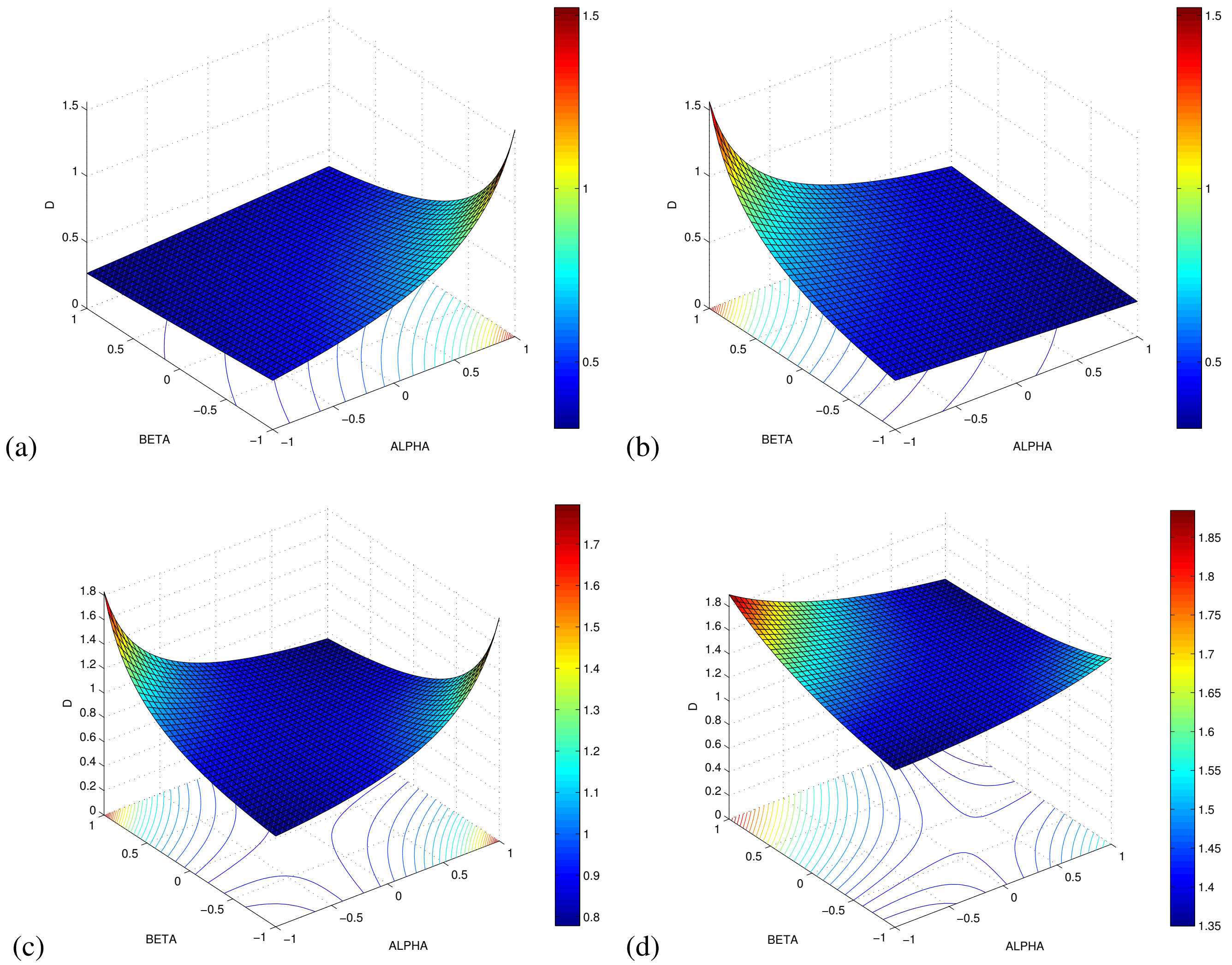


Figure 2. Two-dimensional plots of the AB log-det divergence for different eigenvalues: (**a**) $\lambda = 0.4$, (**b**) $\lambda = 2.5$, (**c**) $\lambda_1 = 0.4, \lambda_2 = 2.5$, (**d**) 10 eigenvalues uniformly randomly distributed in the range $[0.5, 2]$.

In general, the AB log-det divergence is not a metric distance since the triangle inequality is not satisfied for all parameter values. Therefore, we can define the metric distance as the square root of the AB log-det divergence in the special case when $\alpha = \beta$ as follows:

$$d_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q}) = \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})}. \tag{15}$$

This follows from the fact that $D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})$ is symmetric with respect to $\mathbf{P}$ and $\mathbf{Q}$.

Later, we will show that measures defined in this manner lead to many important and well-known divergences and metric distances such as the Logdet Zero divergence, Affine Invariant Riemannian metric (AIRM), and square root of Stein's loss [5,6]. Moreover, new divergences can be generated; specifically, generalized Stein's loss, the Beta-log-det divergence, and extended Hilbert metrics.

From the divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$, a Riemannian metric and a pair of dually coupled affine connections are introduced in the manifold of positive definite matrices. Let $d\mathbf{P}$ be a small deviation of $\mathbf{P}$, which belongs to the tangent space of the manifold at $\mathbf{P}$. Calculating $D_{AB}^{(\alpha,\beta)}(\mathbf{P}+d\mathbf{P}\|\mathbf{P})$ and neglecting higher-order terms yields (see Appendix E)

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}+d\mathbf{P}\|\mathbf{P}) = \frac{1}{2}\,\mathrm{tr}[d\mathbf{P}\,\mathbf{P}^{-1}\,d\mathbf{P}\,\mathbf{P}^{-1}]. \tag{16}$$

This gives a Riemannian metric that is common for all (α,β). Therefore, the Riemannian metric is the same for all AB log-det divergences, although the dual affine connections depend on α and β. The Riemannian metric is also the same as the Fisher information matrix of the manifold of multivariate Gaussian distributions of mean zero and covariance matrix $\mathbf{P}$.

Interestingly, note that the Riemannian metric or geodesic distance is obtained from (3) for $\alpha=\beta=0$:

$$\begin{aligned} d_R(\mathbf{P}\|\mathbf{Q}) &= \sqrt{2\,D_{AB}^{(0,0)}(\mathbf{P}\|\mathbf{Q})} = \sqrt{\mathrm{tr}\log^2(\mathbf{P}\mathbf{Q}^{-1})} && (17)\\ &= \|\log(\mathbf{P}\mathbf{Q}^{-1})\|_F = \|\log(\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})\|_F && (18)\\ &= \sqrt{\sum_{i=1}^{n}\log^2(\lambda_i)}, && (19)\end{aligned}$$

where λ_i are the eigenvalues of $\mathbf{P}\mathbf{Q}^{-1}$.

This is also known as the AIRM. AIRM takes advantage of several important and useful theoretical properties and is probably one of the most widely used (dis)similarity measure for SPD (covariance) matrices [14,15].

For $\alpha=\beta=0.5$ (and for $\alpha=\beta=-0.5$), the recently defined and deeply analyzed S-divergence (JBLD) [6,14,15,17] is obtained:

$$\begin{aligned} D_S(\mathbf{P}\|\mathbf{Q}) &= D_{AB}^{(0.5,0.5)}(\mathbf{P}\|\mathbf{Q}) = 4\log\det\left(\frac{1}{2}\left[(\mathbf{P}\mathbf{Q}^{-1})^{1/2}+(\mathbf{P}\mathbf{Q}^{-1})^{-1/2}\right]\right)\\ &= 4\log\frac{\det(\mathbf{P})^{1/2}\ \det\left(\dfrac{(\mathbf{P}\mathbf{Q}^{-1})^{1/2}+(\mathbf{P}\mathbf{Q}^{-1})^{-1/2}}{2}\right)\ \det(\mathbf{Q})^{1/2}}{\det(\mathbf{P})^{1/2}\det(\mathbf{Q})^{1/2}}\\ &= 4\log\frac{\det\frac{1}{2}(\mathbf{P}+\mathbf{Q})}{\sqrt{\det(\mathbf{P})\det(\mathbf{Q})}}\\ &= 4\left(\log\det\left(\frac{\mathbf{P}+\mathbf{Q}}{2}\right)-\frac{1}{2}\log\det(\mathbf{P}\mathbf{Q})\right) = 4\sum_{i=1}^{n}\log\left(\frac{\lambda_i+1}{2\sqrt{\lambda_i}}\right). && (20)\end{aligned}$$

The S-divergence is not a metric distance. To make it a metric, we take its square root and obtain the LogDet Zero divergence, or Bhattacharyya distance [5,7,18]:

$$\begin{aligned} d_{\mathrm{Bh}}(\mathbf{P}\|\mathbf{Q}) &= \sqrt{D_{AB}^{(0.5,0.5)}(\mathbf{P}\|\mathbf{Q})}\\ &= 2\sqrt{\log\det\left(\frac{\mathbf{P}+\mathbf{Q}}{2}\right)-\frac{1}{2}\log\det(\mathbf{P}\mathbf{Q})}\\ &= 2\sqrt{\log\frac{\det\frac{1}{2}(\mathbf{P}+\mathbf{Q})}{\sqrt{\det(\mathbf{P})\det(\mathbf{Q})}}}\,. && (21)\end{aligned}$$

Moreover, for $\alpha = 0,\ \beta \neq 0$ and $\alpha \neq 0,\ \beta = 0$, we obtain divergences which are generalizations of Stein's loss (called also Burg matrix divergence or simply LogDet divergence):

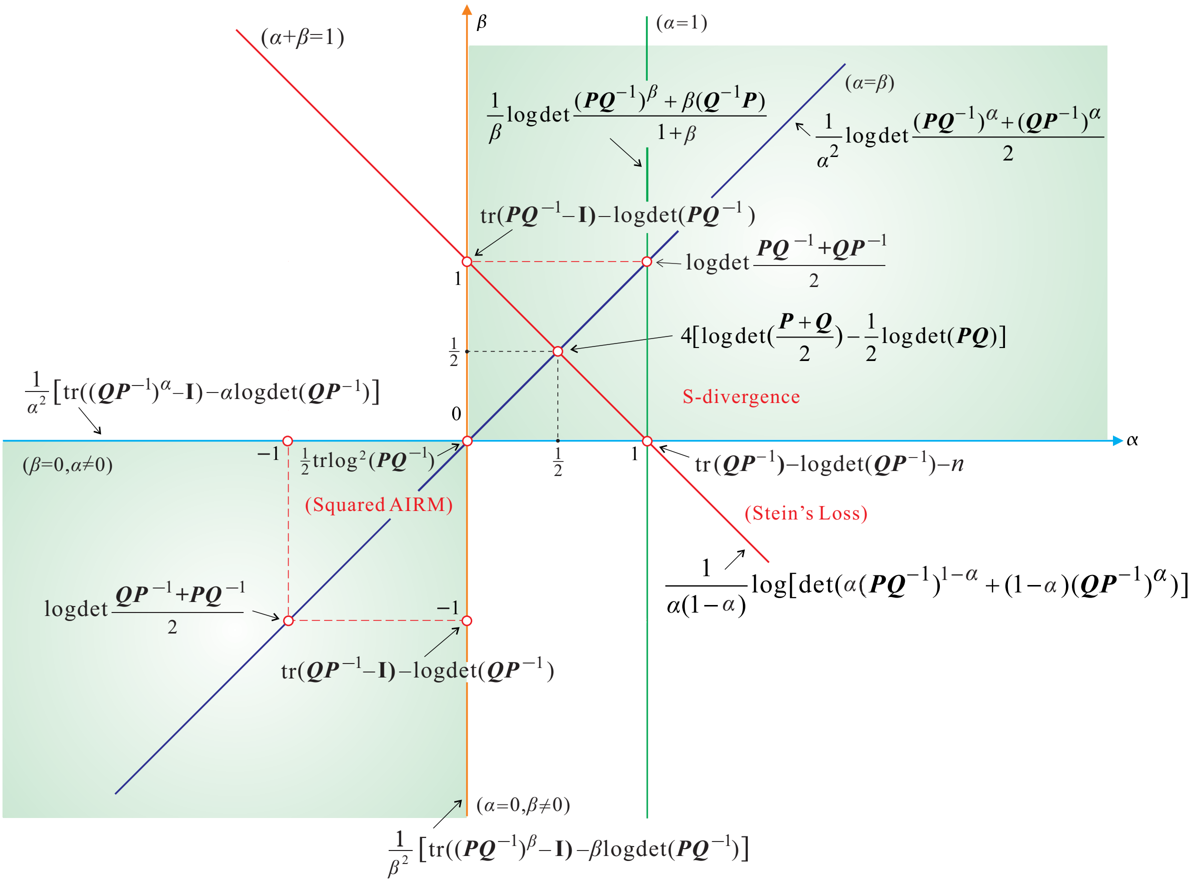
$$D_{AB}^{(0,\beta)}(\mathbf{P}\|\mathbf{Q}) \;=\; \frac{1}{\beta^2}\left[\mathrm{tr}\left((\mathbf{P}\mathbf{Q}^{-1})^{\beta} - \mathbf{I}\right) - \beta \log\det(\mathbf{P}\mathbf{Q}^{-1})\right], \quad \beta \neq 0. \tag{22}$$

$$D_{AB}^{(\alpha,0)}(\mathbf{P}\|\mathbf{Q}) \;=\; \frac{1}{\alpha^2}\left[\mathrm{tr}\left((\mathbf{Q}\mathbf{P}^{-1})^{\alpha} - \mathbf{I}\right) - \alpha \log\det(\mathbf{Q}\mathbf{P}^{-1})\right], \quad \alpha \neq 0 \tag{23}$$

The divergences in (22) and (23) simplify, respectively, to the standard Stein's loss if $\beta = 1$ and to its dual loss if $\alpha = 1$.

3. Special Cases of the AB Log-Det Divergence

We now illustrate how a suitable choice of the (α, β) parameters simplify the AB log-det divergence into other known divergences such as the Alpha- and Beta-log-det divergences [5,11,18,23] (see Figure 3 and Table 1).



$(\alpha+\beta=1)$
β
$(\alpha=1)$
$(\alpha=\beta)$
$\frac{1}{\beta}\mathrm{logdet}\frac{(\boldsymbol{P}\boldsymbol{Q}^{-1})^{\beta}+\beta(\boldsymbol{Q}^{-1}\boldsymbol{P})}{1+\beta}$
$\frac{1}{\alpha^2}\mathrm{logdet}\frac{(\boldsymbol{P}\boldsymbol{Q}^{-1})^{\alpha}+(\boldsymbol{Q}\boldsymbol{P}^{-1})^{\alpha}}{2}$
$\mathrm{tr}(\boldsymbol{P}\boldsymbol{Q}^{-1}-\mathbf{I})-\mathrm{logdet}(\boldsymbol{P}\boldsymbol{Q}^{-1})$
$\mathrm{logdet}\frac{\boldsymbol{P}\boldsymbol{Q}^{-1}+\boldsymbol{Q}\boldsymbol{P}^{-1}}{2}$
1
$4[\mathrm{logdet}(\frac{\boldsymbol{P}+\boldsymbol{Q}}{2})-\frac{1}{2}\mathrm{logdet}(\boldsymbol{P}\boldsymbol{Q})]$
$\frac{1}{2}$
S-divergence
$\frac{1}{\alpha^2}[\mathrm{tr}((\boldsymbol{Q}\boldsymbol{P}^{-1})^{\alpha}-\mathbf{I})-\alpha\mathrm{logdet}(\boldsymbol{Q}\boldsymbol{P}^{-1})]$
0
α
$(\beta=0,\alpha\neq0)$
-1
$\frac{1}{2}\mathrm{trlog}^2(\boldsymbol{P}\boldsymbol{Q}^{-1})$
$\frac{1}{2}$
1
$\mathrm{tr}(\boldsymbol{Q}\boldsymbol{P}^{-1})-\mathrm{logdet}(\boldsymbol{Q}\boldsymbol{P}^{-1})-n$
(Squared AIRM)
(Stein's Loss)
$\frac{1}{\alpha(1-\alpha)}\log[\det(\alpha(\boldsymbol{P}\boldsymbol{Q}^{-1})^{1-\alpha}+(1-\alpha)(\boldsymbol{Q}\boldsymbol{P}^{-1})^{\alpha})]$
$\mathrm{logdet}\frac{\boldsymbol{Q}\boldsymbol{P}^{-1}+\boldsymbol{P}\boldsymbol{Q}^{-1}}{2}$
-1
$\mathrm{tr}(\boldsymbol{Q}\boldsymbol{P}^{-1}-\mathbf{I})-\mathrm{logdet}(\boldsymbol{Q}\boldsymbol{P}^{-1})$
$(\alpha=0,\beta\neq0)$
$\frac{1}{\beta^2}[\mathrm{tr}((\boldsymbol{P}\boldsymbol{Q}^{-1})^{\beta}-\mathbf{I})-\beta\mathrm{logdet}(\boldsymbol{P}\boldsymbol{Q}^{-1})]$

Figure 3. Links between the fundamental, nonsymmetric, AB log-det divergences. On the α-β-plane, important divergences are indicated by points and lines, especially the Stein's loss and its generalization, the AIRM (Riemannian) distance, S-divergence (JBLD), Alpha-log-det divergence $D_A^{(\alpha)}$, and Beta-log-det divergence $D_B^{(\beta)}$.

Table 1. Fundamental Log-det Divergences and Distances

Geodesic Distance (AIRM) $(\alpha = \beta = 0)$ $\tfrac{1}{2}d_R^2(\mathbf{P} \parallel \mathbf{Q}) = \dfrac{1}{2}\,\operatorname{tr}\log^2(\mathbf{P}\mathbf{Q}^{-1}) = \dfrac{1}{2}\sum_{i=1}^{n}\log^2\lambda_i$
S-divergence (Squared Bhattacharyya Distance) $(\alpha = \beta = 0.5)$ $D_S(\mathbf{P} \parallel \mathbf{Q}) = d_{\text{Bh}}^2(\mathbf{P} \parallel \mathbf{Q}) = 4\log\dfrac{\det\frac{\mathbf{P}+\mathbf{Q}}{2}}{(\det\,\mathbf{P}\mathbf{Q})^{\frac{1}{2}}} = 4\sum_{i=1}^{n}\log\dfrac{\lambda_i+1}{2\sqrt{\lambda_i}}$
Power divergence $(\alpha = \beta \neq 0)$ $\dfrac{1}{\alpha^2}\log\det\dfrac{(\mathbf{P}\mathbf{Q}^{-1})^{\alpha}-(\mathbf{P}\mathbf{Q}^{-1})^{-\alpha}}{2} = \dfrac{1}{\alpha^2}\sum\log\dfrac{\lambda_i^{\alpha}+\lambda_i^{-\alpha}}{2}$
Generalized Burg divergence (Stein's Loss) $(\alpha = 0,\ \beta \neq 0)$ $\dfrac{1}{\beta^2}\operatorname{tr}\left[(\mathbf{P}\mathbf{Q}^{-1})^{\beta}-I\right]-\log\det(\mathbf{P}\mathbf{Q}^{-1})^{\beta}\Big] = \dfrac{1}{\beta^2}\Big(\sum_{i=1}^{n}\left(\lambda_i^{\beta}-\log\lambda_i^{\beta}\right)-n\Big)$
Generalized Itakura-Saito log-det divergence $(\alpha = -\beta \neq 0)$ $\dfrac{1}{\alpha^2}\log\dfrac{\det(\mathbf{P}\mathbf{Q}^{-1})^{\alpha}}{\det\,\mathbf{I}+\log(\mathbf{P}\mathbf{Q}^{-1})^{\alpha}} = \dfrac{1}{\alpha^2}\sum_{i=1}^{n}\log\dfrac{\lambda_i^{\alpha}}{1+\log^2\lambda_i^{\alpha}}$
Alpha log-det divergence $(0 < \alpha < 1, \beta = 1-\alpha)$ $D_A^{(\alpha)}(\mathbf{P}\Vert\mathbf{Q}) = \dfrac{1}{\alpha(1-\alpha)}\log\dfrac{\det(\alpha\mathbf{P}+(1-\alpha)\mathbf{Q})}{\det(\mathbf{P}^{\alpha}\,\mathbf{Q}^{1-\alpha})} = \dfrac{1}{\alpha(1-\alpha)}\sum_{i=1}^{n}\log\left(\dfrac{\alpha(\lambda_i-1)+1}{\lambda_i^{\alpha}}\right)$
Beta log-det divergence $(\alpha = 1, \beta \geq 0)$ $D_B^{(\beta)}(\mathbf{P}\Vert\mathbf{Q}) = \dfrac{1}{\beta}\log\det\dfrac{(\mathbf{P}\mathbf{Q}^{-1})^{\beta}+\beta(\mathbf{P}\mathbf{Q}^{-1})}{1+\beta} = \dfrac{1}{\beta}\sum_{i=1}^{n}\log\dfrac{\lambda_i^{\beta}+\beta\lambda_i^{-1}}{1+\beta}$ $D_B^{(\infty)}(\mathbf{P}\Vert\mathbf{Q}) = \sum_{i\in\Omega}\log\lambda_i\,, \quad \Omega = \{i : \lambda_i > 1\}$
Symmetric Jeffrey KL divergence $(\alpha = 1, \beta = 0)$ $D_J(\mathbf{P}\Vert\mathbf{Q}) = \dfrac{1}{2}\operatorname{tr}(\mathbf{P}\mathbf{Q}^{-1}+\mathbf{Q}\mathbf{P}^{-1}-2\mathbf{I}) = \dfrac{1}{2}\sum_{i=1}^{n}\left(\sqrt{\lambda}_i-\dfrac{1}{\sqrt{\lambda}_i}\right)^2$
Generalized Hilbert metrics $D_{CCA}^{(\gamma_2,\gamma_1)}(\mathbf{P}\Vert\mathbf{Q}) = \log\dfrac{M_{\gamma_2}\{\lambda_i\}}{M_{\gamma_1}\{\lambda_i\}}, \quad d_H(\mathbf{P}\Vert\mathbf{Q}) = \log\dfrac{M_{\infty}\{\lambda_i\}}{M_{-\infty}\{\lambda_i\}} = \log\dfrac{\lambda_{max}}{\lambda_{min}}$

When $\alpha+\beta=1$, the AB log-det divergence reduces to the Alpha-log-det divergence [5]:

$$\begin{aligned} D_{AB}^{(\alpha,1-\alpha)}(\mathbf{P}\|\mathbf{Q}) &= D_{A}^{(\alpha)}(\mathbf{P}\|\mathbf{Q}) \qquad (24)\\ &\doteq \begin{cases} \dfrac{1}{\alpha(1-\alpha)}\log\det\left[\alpha(\mathbf{P}\mathbf{Q}^{-1})^{1-\alpha}+(1-\alpha)(\mathbf{Q}\mathbf{P}^{-1})^{\alpha}\right] = \\ \dfrac{1}{\alpha(1-\alpha)}\log\dfrac{\det(\alpha\mathbf{P}+(1-\alpha)\mathbf{Q})}{\det(\mathbf{P}^{\alpha}\,\mathbf{Q}^{1-\alpha})} = \\ \dfrac{1}{\alpha(1-\alpha)}\displaystyle\sum_{i=1}^{n}\log\left(\dfrac{\alpha(\lambda_i-1)+1}{\lambda_i^{\alpha}}\right) & \text{for} \quad 0<\alpha<1,\\ \operatorname{tr}(\mathbf{Q}\mathbf{P}^{-1})-\log\det(\mathbf{Q}\mathbf{P}^{-1})-n=\displaystyle\sum_{i=1}^{n}\left(\lambda_i^{-1}+\log(\lambda_i)\right)-n & \text{for} \quad \alpha=1,\\ \operatorname{tr}(\mathbf{P}\mathbf{Q}^{-1})-\log\det(\mathbf{P}\mathbf{Q}^{-1})-n=\displaystyle\sum_{i=1}^{n}\left(\lambda_i-\log(\lambda_i)\right)-n & \text{for} \quad \alpha=0. \end{cases} \end{aligned}$$

On the other hand, when $\alpha = 1$ and $\beta \geq 0$, the AB log-det divergence reduces to the Beta-log-det divergence:

$$\begin{aligned} D_{AB}^{(1,\beta)}(\mathbf{P}\|\mathbf{Q}) &= D_{B}^{(\beta)}(\mathbf{P}\|\mathbf{Q}) \qquad (25)\\ &\doteq \begin{cases} \dfrac{1}{\beta}\log\det\dfrac{(\mathbf{P}\mathbf{Q}^{-1})^{\beta}+\beta\,(\mathbf{Q}\mathbf{P}^{-1})}{1+\beta} = \dfrac{1}{\beta}\displaystyle\sum_{i=1}^{n}\log\left(\dfrac{\lambda_i^{\beta}+\beta\lambda_i^{-1}}{1+\beta}\right) & \text{for} \quad \beta>0,\\ \operatorname{tr}(\mathbf{Q}\mathbf{P}^{-1}-\mathbf{I})-\log\det(\mathbf{Q}\mathbf{P}^{-1}) = \displaystyle\sum_{i=1}^{n}\left(\lambda_i^{-1}+\log(\lambda_i)\right)-n & \text{for} \quad \beta=0,\\ \log\dfrac{\det(\mathbf{P}\mathbf{Q}^{-1})}{\det(\mathbf{I}+\log(\mathbf{P}\mathbf{Q}^{-1}))} = \displaystyle\sum_{i=1}^{n}\log\dfrac{\lambda_i}{1+\log(\lambda_i)} & \text{for} \quad \beta=-1,\ \lambda_i>e^{-1}\forall i. \end{cases} \end{aligned}$$

Note that $\det(\mathbf{I}+\log(\mathbf{P}\mathbf{Q}^{-1}) = \prod_{i=1}^{n}[1+\log(\lambda_i)]$, and the Beta-log-det divergence is well defined for $\beta=-1$ and if all the eigenvalues are larger than $\lambda_i > e^{-1} \approx 0.367$ ($e \approx 2.72$).

It is interesting to note that the Beta-log-det divergence for $\beta\to\infty$ leads to a new divergence that is robust with respect to noise. This new divergence is given by

$$\lim_{\beta\to\infty} D_{B}^{(\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{B}^{(\infty)}(\mathbf{P}\|\mathbf{Q}) = \log(\prod_{i=1}^{k}\lambda_i) \ \text{ for all } \ \lambda_i \geq 1. \qquad (26)$$

This can be easily shown by applying the L'Hôpital's formula. Assuming that the set $\Omega=\{i:\lambda_i>1\}$ gathers the indices of those eigenvalues greater than one, we can more formally express this divergence as

$$D_{B}^{(\infty)}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \sum_{i\in\Omega}\log\lambda_i & \text{for } \Omega\neq\phi,\\ 0 & \text{for } \Omega=\phi. \end{cases} \qquad (27)$$

The Alpha-log-det divergence gives the standard Stein's losses (Burg matrix divergences) for $\alpha=1$ and $\alpha=0$, and the Beta-log-det divergence is equivalent to Stein's loss for $\beta=0$.

Another important class of divergences is Power log-det divergences for any $\alpha = \beta \in \mathbb{R}$:

$$D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q}) = D_{P}^{(\alpha)}(\mathbf{P}\|\mathbf{Q}) \tag{28}$$

$$\doteq \begin{cases} \dfrac{1}{\alpha^2}\log\det\dfrac{(\mathbf{P}\mathbf{Q}^{-1})^{\alpha}+(\mathbf{P}\mathbf{Q}^{-1})^{-\alpha}}{2} = \dfrac{1}{\alpha^2}\displaystyle\sum_{i=1}^{n}\log\dfrac{\lambda_i^{\alpha}+\lambda_i^{-\alpha}}{2} & \text{for} \quad \alpha \neq 0, \\ \\ \dfrac{1}{2}\operatorname{tr}\log^2(\mathbf{P}\mathbf{Q}^{-1}) = \dfrac{1}{2}\operatorname{tr}\log^2(\mathbf{Q}\mathbf{P}^{-1}) = \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\log^2(\lambda_i) & \text{for} \quad \alpha = 0. \end{cases}$$

4. Properties of the AB Log-Det Divergence

The AB log-det divergence has several important and useful theoretical properties for SPD matrices.

1. Nonnegativity; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \geq 0, \qquad \forall \alpha,\beta \in \mathbb{R}. \tag{29}$$

2. Identity of indiscernibles (see Theorems 1 and 2); given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = 0 \ \text{ if and only if } \ \mathbf{P} = \mathbf{Q}. \tag{30}$$

3. Continuity and smoothness of $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ as a function of $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$, including the singular cases when $\alpha = 0$ or $\beta = 0$, and when $\alpha = -\beta$ (see Figure 2).

4. The divergence can be expressed in terms of the diagonal matrix $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ with the eigenvalues of $\mathbf{P}\mathbf{Q}^{-1}$, in the form

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\mathbf{\Lambda}\|\mathbf{I}). \tag{31}$$

5. Scaling invariance; given by

$$D_{AB}^{(\alpha,\beta)}(c\mathbf{P}\|c\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}), \tag{32}$$

   for any $c > 0$.

6. Relative invariance for scale transformation: For given α and β and nonzero scaling factor $\omega \neq 0$, we have

$$D_{AB}^{(\omega\,\alpha,\,\omega\,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\omega^2} D_{AB}^{(\alpha,\beta)}((\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})^{\omega}\|\mathbf{I})\,. \tag{33}$$

7. Dual-invariance under inversion (for $\omega = -1$); given by

$$D_{AB}^{(-\alpha,-\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\mathbf{P}^{-1}\|\mathbf{Q}^{-1})\,. \tag{34}$$

8. Dual symmetry; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\beta,\alpha)}(\mathbf{Q}\|\mathbf{P})\,. \tag{35}$$

9. Affine invariance (invariance under congruence transformations); given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{A}\mathbf{P}\mathbf{A}^T\|\mathbf{A}\mathbf{Q}\mathbf{A}^T) \;=\; D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}), \tag{36}$$

for any nonsingular matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$

10. Divergence lower-bound; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{X}^T\mathbf{P}\mathbf{X}\|\mathbf{X}^T\mathbf{Q}\mathbf{X}) \;\leq\; D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}), \tag{37}$$

for any full-column rank matrix $\mathbf{X} \in \mathbb{R}^{n\times m}$ with $n \leq m$.

11. Scaling invariance under the Kronecker product; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{Z}\otimes\mathbf{P}\|\mathbf{Z}\otimes\mathbf{Q}) = n\, D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})\;, \tag{38}$$

for any symmetric and positive definite matrix $\mathbf{Z}$ of rank n.

12. Double Sided Orthogonal Procrustes property. Consider an orthogonal matrix $\mathbf{\Omega} \in \mathcal{O}(n)$ and two symmetric positive definite matrices $\mathbf{P}$ and $\mathbf{Q}$, with respective eigenvalue matrices $\mathbf{\Lambda_P}$ and $\mathbf{\Lambda_Q}$ which elements are sorted in descending order. The AB log-det divergence between $\mathbf{\Omega}^T\mathbf{P}\mathbf{\Omega}$ and $\mathbf{Q}$ is globally minimized when their eigenspaces are aligned, i.e.,

$$\min_{\mathbf{\Omega}\in\mathcal{O}(n)} D_{AB}^{(\alpha,\beta)}(\mathbf{\Omega}^T\mathbf{P}\mathbf{\Omega}\|\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\mathbf{\Lambda_P}\|\mathbf{\Lambda_Q}). \tag{39}$$

13. Triangle Inequality-Metric Distance Condition, for $\alpha = \beta \in \mathbb{R}$. The previous property implies the validity of the triangle inequality for arbitrary positive definite matrices, i.e.,

$$\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})} \leq \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Z})} + \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{Z}\|\mathbf{Q})}\;. \tag{40}$$

The proof of this property exploits the metric characterization of the square root of the S-divergence proposed first by S. Sra in [6,17] for arbitrary SPD matrices.

Several of these properties have been already proved for the specific cases of α and β that lead to the S-divergence $(\alpha, \beta = 1/2)$ [6], the Alpha log-det divergence $(0 \leq \alpha \leq 1, \beta = 1-\alpha)$ [5] and the Riemannian metric $(\alpha, \beta = 0)$ [28, Chapter 6]. We refer the reader to Appendix F for their proofs when $\alpha, \beta \in \mathbb{R}$.

5. Symmetrized AB Log-Det Divergences

The basic AB log-det divergence is asymmetric; that is, $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\,\|\,\mathbf{Q}) \neq D_{AB}^{(\alpha,\beta)}(\mathbf{Q}\,\|\,\mathbf{P})$, except the spacial case of $\alpha = \beta$).

In general, there are several ways to symmetrize a divergence; for example, Type-1,

$$D_{ABS1}^{(\alpha,\beta)}(\mathbf{P}\,\|\,\mathbf{Q}) = \frac{1}{2}\left[D_{AB}^{(\alpha,\beta)}(\mathbf{P}\,\|\,\mathbf{Q}) + D_{AB}^{(\alpha,\beta)}(\mathbf{Q}\,\|\,\mathbf{P})\right], \tag{41}$$

and Type-2, based on the Jensen-Shannon symmetrization (which is too complex for log-det divergences),

$$D^{(\alpha,\beta)}_{ABS2}(\mathbf{P} \,\|\, \mathbf{Q}) = \frac{1}{2}\left[D^{(\alpha,\beta)}_{AB}\left(\mathbf{P} \,\|\, \frac{\mathbf{P}+\mathbf{Q}}{2}\right) + D^{(\alpha,\beta)}_{AB}\left(\mathbf{Q} \,\|\, \frac{\mathbf{P}+\mathbf{Q}}{2}\right)\right]. \tag{42}$$

The Type-1 symmetric AB log-det divergence is defined as

$$D^{(\alpha,\beta)}_{ABS1}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \dfrac{1}{2\alpha\beta}\log\det\left[\mathbf{I} + \dfrac{\alpha\beta}{(\alpha+\beta)^2}\left((\mathbf{P}\mathbf{Q}^{-1})^{\alpha+\beta} + (\mathbf{Q}\mathbf{P}^{-1})^{\alpha+\beta} - 2\mathbf{I}\right)\right] & \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } \alpha\beta > 0, & \\ \dfrac{1}{2\alpha^2}\left[\operatorname{tr}\left((\mathbf{P}\mathbf{Q}^{-1})^{\alpha} + (\mathbf{Q}\mathbf{P}^{-1})^{\alpha} - 2\mathbf{I}\right)\right] & \text{for } \alpha \neq 0,\ \beta = 0, \\ \dfrac{1}{2\beta^2}\left[\operatorname{tr}\left((\mathbf{P}\mathbf{Q}^{-1})^{\beta} + (\mathbf{Q}\mathbf{P}^{-1})^{\beta} - 2\mathbf{I}\right)\right] & \text{for } \alpha = 0,\ \beta \neq 0, \\ \dfrac{1}{2\alpha^2}\operatorname{tr}\log(\mathbf{I} - \log^2(\mathbf{P}\mathbf{Q}^{-1})^{\alpha})^{-1} & \text{for } \alpha = -\beta \neq 0, \\ \dfrac{1}{2}\operatorname{tr}\log^2(\mathbf{P}\mathbf{Q}^{-1}) = \dfrac{1}{2}||\log(\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})||^2_F & \text{for } \alpha,\ \beta = 0. \end{cases} \tag{43}$$

Equivalently, this can be expressed by the eigenvalues of $\mathbf{P}\mathbf{Q}^{-1}$ in the form

$$D^{(\alpha,\beta)}_{ABS1}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \dfrac{1}{2\alpha\beta}\displaystyle\sum_{i=1}^{n}\log\left(1 + \dfrac{\alpha\beta}{(\alpha+\beta)^2}(\lambda_i^{\frac{\alpha+\beta}{2}} - \lambda_i^{-\frac{\alpha+\beta}{2}})^2\right) & \text{for } \alpha\beta > 0, \\ \dfrac{1}{2\alpha^2}\displaystyle\sum_{i=1}^{n}\left(\lambda_i^{\alpha} + \lambda_i^{-\alpha} - 2\right) = \dfrac{1}{2\alpha^2}\displaystyle\sum_{i=1}^{n}(\lambda_i^{\frac{\alpha}{2}} - \lambda_i^{-\frac{\alpha}{2}})^2 & \text{for } \alpha \neq 0,\ \beta = 0, \\ \dfrac{1}{2\beta^2}\displaystyle\sum_{i=1}^{n}\left(\lambda_i^{\beta} + \lambda_i^{-\beta} - 2\right) = \dfrac{1}{2\beta^2}\displaystyle\sum_{i=1}^{n}(\lambda_i^{\frac{\beta}{2}} - \lambda_i^{-\frac{\beta}{2}})^2 & \text{for } \alpha = 0,\ \beta \neq 0, \\ \dfrac{1}{2\alpha^2}\displaystyle\sum_{i=1}^{n}\log\dfrac{1}{1 - \log^2(\lambda_i^{\alpha})} & \text{for } \alpha = -\beta \neq 0, \\ \dfrac{1}{2}\displaystyle\sum_{i=1}^{n}\log^2(\lambda_i) & \text{for } \alpha,\ \beta = 0. \end{cases} \tag{44}$$

We consider several well-known symmetric log-det divergences (see Figure 4); in particular, we consider the following:

(1) For $\alpha = \beta = \pm 0.5$, we obtain the S-divergence or JBLD divergence (20).

(2) For $\alpha = \beta = 0$, we obtain the square of the AIRM (Riemannian metric) (19).

(3) For $\alpha = 0$ and $\beta = \pm 1$ or for $\beta = 0$ and $\alpha = \pm 1$, we obtain the KLDM (symmetrized KL Density Metric), also known as the symmetric Stein's loss or Jeffreys KL divergence [3]:

$$\begin{aligned} D_J(\mathbf{P}\|\mathbf{Q}) &= \frac{1}{2}\,\mathrm{tr}\left(\mathbf{P}\mathbf{Q}^{-1} + \mathbf{Q}\mathbf{P}^{-1} - 2\,\mathbf{I}\right) \\ &= \frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{\lambda_i} - \frac{1}{\sqrt{\lambda_i}}\right)^2 . \end{aligned} \tag{45}$$

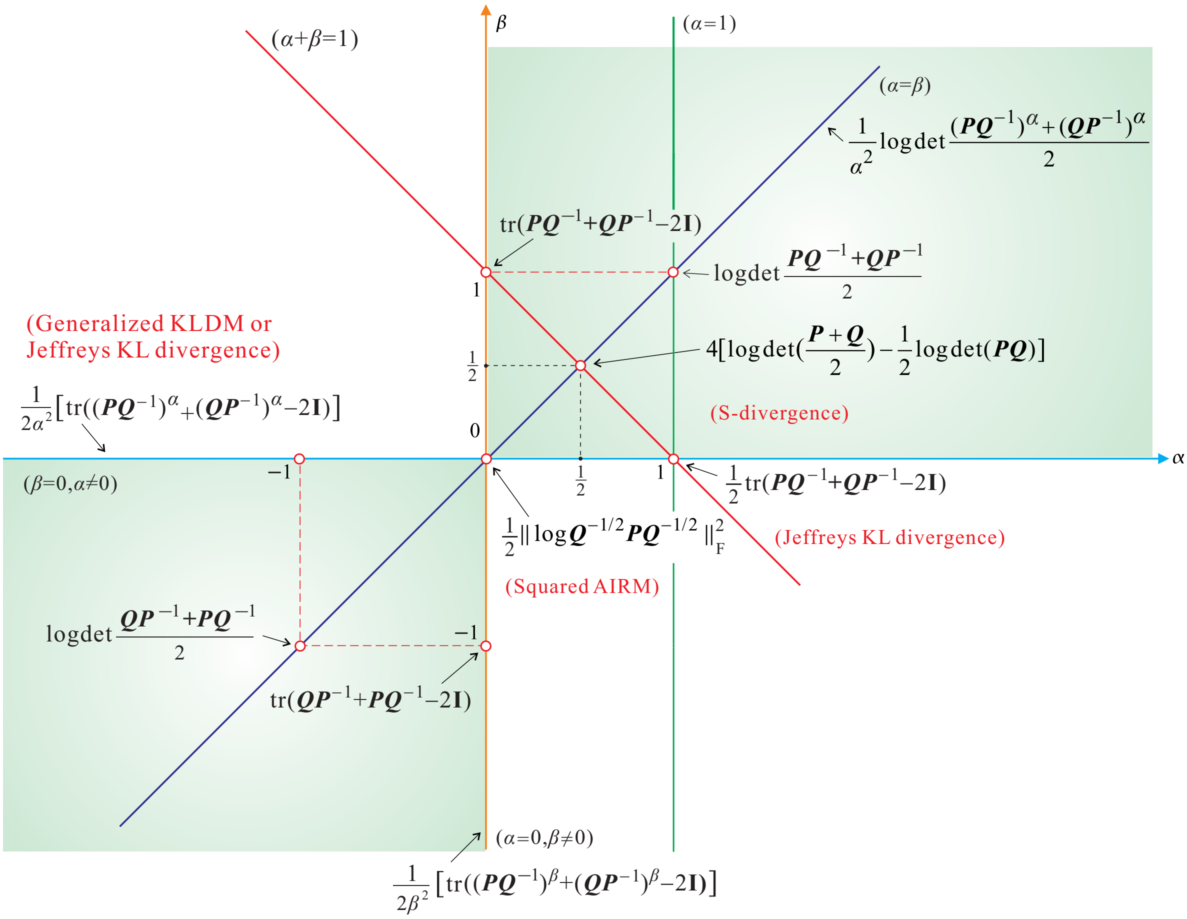

$(\alpha+\beta=1)$
β
$(\alpha=1)$
$(\alpha=\beta)$
$\frac{1}{\alpha^2}\log\det\frac{(PQ^{-1})^{\alpha}+(QP^{-1})^{\alpha}}{2}$
$\mathrm{tr}(PQ^{-1}+QP^{-1}-2\mathbf{I})$
$\log\det\frac{PQ^{-1}+QP^{-1}}{2}$
1
(Generalized KLDM or Jeffreys KL divergence)
$\frac{1}{2}$
$4[\log\det(\frac{P+Q}{2})-\frac{1}{2}\log\det(PQ)]$
(S-divergence)
$\frac{1}{2\alpha^2}[\mathrm{tr}((PQ^{-1})^{\alpha}+(QP^{-1})^{\alpha}-2\mathbf{I})]$
0
α
$(\beta=0,\alpha\neq0)$
-1
$\frac{1}{2}$
1
$\frac{1}{2}\mathrm{tr}(PQ^{-1}+QP^{-1}-2\mathbf{I})$
$\frac{1}{2}\|\log Q^{-1/2}PQ^{-1/2}\|_F^2$
(Jeffreys KL divergence)
(Squared AIRM)
$\log\det\frac{QP^{-1}+PQ^{-1}}{2}$
-1
$\mathrm{tr}(QP^{-1}+PQ^{-1}-2\mathbf{I})$
$(\alpha=0,\beta\neq0)$
$\frac{1}{2\beta^2}[\mathrm{tr}((PQ^{-1})^{\beta}+(QP^{-1})^{\beta}-2\mathbf{I})]$

Figure 4. Links between the fundamental symmetric, AB log-det divergences. On the (α, β)-plane, the special cases of particular divergences are indicated by points (Jeffreys KL divergence (KLDM) or symmetric Stein's loss and its generalization, S-divergence (JBLD), and the Power log-det divergence.

One important potential application of the AB log-det divergence is to generate conditionally positive definite kernels, which are widely applied to classification and clustering. For a specific set of parameters, the AB log-det divergence gives rise to a Hilbert space embedding in the form of a Radial Basis Function (RBF) kernel [22]; more specifically, the AB log-det kernel is defined by

$$\begin{aligned} K_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) &= \exp\left(-\gamma D_{ABS1}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})\right) \\ &= \left(\det\left[\mathbf{I} + \frac{\alpha\beta}{(\alpha+\beta)^2}\left((\mathbf{P}\mathbf{Q}^{-1})^{\alpha+\beta} + (\mathbf{Q}\mathbf{P}^{-1})^{\alpha+\beta} - 2\mathbf{I}\right)\right]\right)^{-\frac{\gamma}{2\alpha\beta}} \end{aligned} \tag{46}$$

for some selected values of $\gamma > 0$ and $\alpha, \beta > 0$ or $\alpha, \beta < 0$ that can make the kernel positive definite.

6. Similarity Measures for Semidefinite Covariance Matrices in Reproducing Kernel Hilbert Spaces

There are many practical applications for which the underlying covariance matrices are symmetric but only positive semidefinite, i.e., their columns do not span the whole space. For instance, in classification problems, assume two classes and a set of observation vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ and $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$ in $\mathbb{R}^m$ for each class, then we may wish to find a principled way to evaluate the ensemble similarity of the data from their sample similarity. The problem of the modeling of similarity between two ensembles was studied by Zhou and Chellappa in [32]. For this purpose, they proposed several probabilistic divergence measures between positive semidefinite covariance matrices in a Reproducing kernel Hilbert space (RKHS) of finite dimensionality. Their strategy was later extended for image classification problems [33] and formalized for the Log-Hilbert-Schmidt metric between infinite-dimensional RKHS covariance operators [34].

In this section, we propose the unifying framework of the AB log-det divergences to reinterpret and extend the similarity measures obtained in [32,33] for semidefinite covariance matrices in the finite-dimensional RKHS.

We shall assume that the nonlinear functions $\Phi_x : \mathbb{R}^m \to \mathbb{R}^n$ and $\Phi_y : \mathbb{R}^m \to \mathbb{R}^n$ (where $n > m$) respectively map the data from each of the classes into their higher dimensional feature spaces. We implicitly define the feature matrices as

$$\mathbf{\Phi}_{\boldsymbol{x}} = [\Phi_x(\boldsymbol{x}_1), \ldots, \Phi_x(\boldsymbol{x}_T)], \qquad \mathbf{\Phi}_{\boldsymbol{y}} = [\Phi_y(\boldsymbol{y}_1), \ldots, \Phi_y(\boldsymbol{y}_T)], \tag{47}$$

and the sample covariance matrices of the observations in the feature space as: $\mathbf{C}_{\boldsymbol{x}} = \mathbf{\Phi}_{\boldsymbol{x}} \mathbf{J} \mathbf{\Phi}_{\boldsymbol{x}}^T / T \in \mathbb{R}^{n \times n}$ and $\mathbf{C}_{\boldsymbol{y}} = \Phi_{\boldsymbol{y}} \mathbf{J} \Phi_{\boldsymbol{y}}^T / T \in \mathbb{R}^{n \times n}$, where $\mathbf{J} = \mathbf{I}_T - \frac{1}{T}\mathbf{1}\mathbf{1}^T$ denotes the $T \times T$ centering matrix.

In practice, it is common to consider low-rank approximations of sample covariance matrices. For a given basis $\mathbf{V}_{\boldsymbol{x}} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r) \in \mathbb{R}^{T \times r}$ of the principal subspace of $\mathbf{J}\mathbf{\Phi}_{\boldsymbol{x}}^T \mathbf{\Phi}_{\boldsymbol{x}} \mathbf{J}$, we can define the projection matrix $\mathbf{\Pi}_{\boldsymbol{x}} = \mathbf{V}_{\boldsymbol{x}} \mathbf{V}_{\boldsymbol{x}}^T$ and redefine the covariance matrices as

$$\mathbf{C}_{\boldsymbol{x}} = \frac{1}{T} \mathbf{\Phi}_{\boldsymbol{x}} \, \mathbf{V}_{\boldsymbol{x}} \mathbf{V}_{\boldsymbol{x}}^T \, \mathbf{\Phi}_{\boldsymbol{x}}^T \qquad \text{and} \qquad \mathbf{C}_{\boldsymbol{y}} = \frac{1}{T} \mathbf{\Phi}_{\boldsymbol{y}} \, \mathbf{V}_{\boldsymbol{y}} \mathbf{V}_{\boldsymbol{y}}^T \, \mathbf{\Phi}_{\boldsymbol{y}}^T. \tag{48}$$

Assuming the Gaussianity of the data in the feature space, the mean vector and covariance matrix are sufficient statistics and a natural measure of dissimilarity between $\mathbf{\Phi}_{\boldsymbol{x}}$ and $\mathbf{\Phi}_{\boldsymbol{y}}$ should be a function of the first and second order statistics of the features. Furthermore, in most practical problems the mean value should be ignored due to robustness considerations, and then the comparison reduces to the evaluation of a suitable dissimilarity measure between $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$.

The dimensionality of the feature space n is typically much larger than r, so the rank of the covariance matrices in (48) will be $r \ll n$ and, therefore, both matrices are positive semidefinite. The AB log-det divergence is infinite when the range spaces of the covariance matrices $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$ differ. This property is useful in applications which require an automatic constraint in the range of the estimates [22], but it will prohibit the practical use of the comparison when the ranges of the covariance matrices differ. The next subsections present two different strategies to address this challenging problem.

6.1. Measuring the Dissimilarity with a Divergence Lower-Bound

One possible strategy is to use dissimilarity measures which ignore the contribution to the divergence caused by the rank deficiency of the covariance matrices. This is useful when performing one comparison of the covariances matrices after applying a congruence transformation that aligns their range spaces, and can be implemented by retaining only the finite and non-zero eigenvalues of the matrix pencil $(\mathbf{C}_{\boldsymbol{x}}, \mathbf{C}_{\boldsymbol{y}})$.

Let $\mathbf{I}_r$ denote the identity matrix of size r and $(\cdot)^+$ the Moore-Penrose pseudoinverse operator. Consider the eigenvalue decomposition of the symmetric matrix

$$(\mathbf{C}_{\boldsymbol{y}}^+)^{\frac{1}{2}}\mathbf{C}_{\boldsymbol{x}}(\mathbf{C}_{\boldsymbol{y}}^+)^{\frac{1}{2}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \tag{49}$$

where $\mathbf{U}$ is a semi-orthogonal matrix for which the columns are the eigenvectors associated with the positive eigenvalues of the matrix pencil and

$$\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_r) \;\; \equiv \;\; \operatorname{diag} Eig_+\{(\mathbf{C}_{\boldsymbol{y}}^+)^{\frac{1}{2}}\mathbf{C}_{\boldsymbol{x}}(\mathbf{C}_{\boldsymbol{y}}^+)^{\frac{1}{2}})\}. \tag{50}$$

is a diagonal matrix with the eigenvalues sorted in a descending order.

Note that the tall matrix $\mathbf{W} = (\mathbf{C}_{\boldsymbol{y}}^+)^{\frac{1}{2}}\mathbf{U} \in \mathbb{R}^{n\times r}$ diagonalizes the covariance matrices of the two classes

$$\mathbf{W}^T\mathbf{C}_{\boldsymbol{x}}\mathbf{W} \;\; = \;\; \boldsymbol{\Lambda} \tag{51}$$

$$\mathbf{W}^T\mathbf{C}_{\boldsymbol{y}}\mathbf{W} \;\; = \;\; \mathbf{I}_r \tag{52}$$

and compress them to a common range space. The compression automatically discards the singular and infinite eigenvalues of the matrix pencil $(\mathbf{C}_{\boldsymbol{x}}, \mathbf{C}_{\boldsymbol{y}})$, while it retains the finite and positive eigenvalues. In this way, the following dissimilarity measures can be obtained:

$$L_{AB}^{(\alpha,\beta)}(\mathbf{C}_{\boldsymbol{x}}, \mathbf{C}_{\boldsymbol{y}}) \;\; \equiv \;\; D_{AB}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{C}_{\boldsymbol{x}}\mathbf{W}\|\mathbf{W}^T\mathbf{C}_{\boldsymbol{y}}\mathbf{W}) = D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Lambda}\|\mathbf{I}_r), \tag{53}$$

$$L_{ABS1}^{(\alpha,\beta)}(\mathbf{C}_{\boldsymbol{x}}, \mathbf{C}_{\boldsymbol{y}}) \;\; \equiv \;\; D_{ABS1}^{(\alpha,\beta)}(\mathbf{W}^T\mathbf{C}_{\boldsymbol{x}}\mathbf{W}\|\mathbf{W}^T\mathbf{C}_{\boldsymbol{y}}\mathbf{W}) = D_{ABS1}^{(\alpha,\beta)}(\boldsymbol{\Lambda}\|\mathbf{I}_r). \tag{54}$$

Note, however, that these measures should not be understood as a strict comparison of the original covariance matrices, but rather as an indirect comparison through their respective compressed versions $\mathbf{W}^T\mathbf{C}_{\boldsymbol{x}}\mathbf{W}$ and $\mathbf{W}^T\mathbf{C}_{\boldsymbol{y}}\mathbf{W}$.

With the help of the kernel trick, the next lemma shows that the evaluation of the dissimilarity measures $L_{AB}^{(\alpha,\beta)}(\mathbf{C}_{\boldsymbol{x}}, \mathbf{C}_{\boldsymbol{y}})$ and $L_{ABS1}^{(\alpha,\beta)}(\mathbf{C}_{\boldsymbol{x}}, \mathbf{C}_{\boldsymbol{y}})$, does not require the explicit computation of the covariance matrices or of the feature vectors.

Lemma 1. *Given the Gram matrix or kernel matrix of the input vectors*

$$\begin{pmatrix} \mathbf{K}_{\boldsymbol{xx}} & \mathbf{K}_{\boldsymbol{xy}} \\ \mathbf{K}_{\boldsymbol{yx}} & \mathbf{K}_{\boldsymbol{yy}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Phi}_{\boldsymbol{x}}^T\boldsymbol{\Phi}_{\boldsymbol{x}} & \boldsymbol{\Phi}_{\boldsymbol{x}}^T\boldsymbol{\Phi}_{\boldsymbol{y}} \\ \boldsymbol{\Phi}_{\boldsymbol{y}}^T\boldsymbol{\Phi}_{\boldsymbol{x}} & \boldsymbol{\Phi}_{\boldsymbol{y}}^T\boldsymbol{\Phi}_{\boldsymbol{y}} \end{pmatrix} \tag{55}$$

and the matrices $\mathbf{V}_{\boldsymbol{x}}$ *and* $\mathbf{V}_{\boldsymbol{y}}$ *which respectively span the principal subspaces of* $\mathbf{K}_{\boldsymbol{xx}}$ *and* $\mathbf{K}_{\boldsymbol{yy}}$*, the positive and finite eigenvalues of the matrix pencil can be expressed by*

$$\boldsymbol{\Lambda} \;\; = \;\; \operatorname{diag} Eig_+\left\{(\mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{xy}}\mathbf{K}_{\boldsymbol{yy}}^{-1}\mathbf{V}_{\boldsymbol{y}})(\mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{xy}}\mathbf{K}_{\boldsymbol{yy}}^{-1}\mathbf{V}_{\boldsymbol{y}})^T\right\}. \tag{56}$$

Proof. The proof of the lemma relies on the property that for any pair of $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, the non-zero eigenvalues of $\mathbf{A}\mathbf{B}^T$ and of $\mathbf{B}^T\mathbf{A}$ are the same (see [30, pag. 11]). Then, there is an equality between the following matrices of positive eigenvalues

$$\mathbf{\Lambda} \equiv \mathrm{diag}\, Eig_+ \left\{ (\mathbf{C}_{\boldsymbol{y}}^+)^{\frac{1}{2}} \mathbf{C}_{\boldsymbol{x}} (\mathbf{C}_{\boldsymbol{y}}^+)^{\frac{1}{2}}) \right\} = \mathrm{diag}\, Eig_+ \left\{ \mathbf{C}_{\boldsymbol{x}} \mathbf{C}_{\boldsymbol{y}}^+ \right\}. \tag{57}$$

Taking into account the structure of the covariance matrices in (48), such eigenvalues can be explicitly obtained in terms of the kernel matrices

$$\begin{aligned} Eig_+ \left\{ \mathbf{C}_{\boldsymbol{x}} \mathbf{C}_{\boldsymbol{y}}^+ \right\} &= Eig_+ \left\{ (\mathbf{\Phi}_{\boldsymbol{x}} \mathbf{V}_{\boldsymbol{x}} \mathbf{V}_{\boldsymbol{x}}^T \mathbf{\Phi}_{\boldsymbol{x}}^T)((\mathbf{\Phi}_{\boldsymbol{y}}^+)^T \mathbf{V}_{\boldsymbol{y}} \mathbf{V}_{\boldsymbol{y}}^T \mathbf{\Phi}_{\boldsymbol{y}}^+) \right\} & (58) \\ &= Eig_+ \left\{ (\mathbf{V}_{\boldsymbol{x}}^T \mathbf{\Phi}_{\boldsymbol{x}}^T (\mathbf{\Phi}_{\boldsymbol{y}}^T)^+ \mathbf{V}_{\boldsymbol{y}} \mathbf{V}_{\boldsymbol{y}}^T \mathbf{\Phi}_{\boldsymbol{y}}^+)(\mathbf{\Phi}_{\boldsymbol{x}} \mathbf{V}_{\boldsymbol{x}}) \right\} & (59) \\ &= Eig_+ \left\{ (\mathbf{V}_{\boldsymbol{x}}^T \mathbf{K}_{\boldsymbol{xy}} \mathbf{K}_{\boldsymbol{yy}}^{-1} \mathbf{V}_{\boldsymbol{y}})(\mathbf{V}_{\boldsymbol{x}}^T \mathbf{K}_{\boldsymbol{xy}} \mathbf{K}_{\boldsymbol{yy}}^{-1} \mathbf{V}_{\boldsymbol{y}})^T \right\}. & (60) \end{aligned}$$

□

6.2. Similarity Measures Between Regularized Covariance Descriptors

Several authors consider a completely different strategy, which consists in the regularization of the original covariance matrices [32–34]. This way the null the eigenvalues of the covariances $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$ are replaced by a small positive constant $\rho > 0$, to obtain the "regularized" positive definite matrices $\tilde{\mathbf{C}}_{\boldsymbol{x}}$ and $\tilde{\mathbf{C}}_{\boldsymbol{y}}$, respectively. The modification can be illustrated by comparing the eigendecompositions

$$\begin{aligned} \mathbf{C}_{\boldsymbol{x}} &= (\mathbf{U}_{\boldsymbol{x}} | \mathbf{U}_{\boldsymbol{x}}^{\perp}) \begin{pmatrix} \mathbf{\Lambda}_{\boldsymbol{x}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{U}_{\boldsymbol{x}} | \mathbf{U}_{\boldsymbol{x}}^{\perp})^T = \mathbf{U}_{\boldsymbol{x}} \mathbf{\Lambda}_{\boldsymbol{x}} \mathbf{U}_{\boldsymbol{x}}^T & (61) \\ \downarrow & & (62) \\ \tilde{\mathbf{C}}_{\boldsymbol{x}} &= (\mathbf{U}_{\boldsymbol{x}} | \mathbf{U}_{\boldsymbol{x}}^{\perp}) \begin{pmatrix} \mathbf{\Lambda}_{\boldsymbol{x}} & \mathbf{0} \\ \mathbf{0} & \rho\, \mathbf{I}_{n-r} \end{pmatrix} (\mathbf{U}_{\boldsymbol{x}} | \mathbf{U}_{\boldsymbol{x}}^{\perp})^T = \mathbf{C}_{\boldsymbol{x}} + \rho\, \mathbf{U}_{\boldsymbol{x}}^{\perp} (\mathbf{U}_{\boldsymbol{x}}^{\perp})^T. & (63) \end{aligned}$$

Then, the dissimilarity measure of the data in the feature space can be obtained just by measuring a divergence between the SPD matrices $\tilde{\mathbf{C}}_{\boldsymbol{x}}$ and $\tilde{\mathbf{C}}_{\boldsymbol{y}}$. Again, the idea is to compute the value of the divergence without requiring the evaluation of the feature vectors but by using the available kernels.

Using the properties of the trace and the determinants, a practical formula for the log-det Alpha-divergence has been obtained in [32,33] for $0 < \alpha < 1$. The resulting expression

$$D_{AB}^{(\alpha,1-\alpha)}(\tilde{\mathbf{C}}_{\boldsymbol{x}} \,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) = \frac{1}{\alpha(1-\alpha)} \log\det(\mathbf{I}_{2r} + \rho^{-1}\mathbf{H}) - \frac{1}{(1-\alpha)} \log\det(\rho^{-1}\mathbf{\Lambda}_{\boldsymbol{x}}) - \frac{1}{\alpha} \log\det(\rho^{-1}\mathbf{\Lambda}_{\boldsymbol{y}})$$

is a function of the principal eigenvalues of the kernels

$$\mathbf{\Lambda}_{\boldsymbol{x}} = \mathbf{V}_{\boldsymbol{x}}^T \mathbf{K}_{\boldsymbol{xx}} \mathbf{V}_{\boldsymbol{x}}, \qquad \mathbf{\Lambda}_{\boldsymbol{y}} = \mathbf{V}_{\boldsymbol{x}}^T \mathbf{K}_{\boldsymbol{yy}} \mathbf{V}_{\boldsymbol{y}}, \tag{64}$$

and the matrix

$$\mathbf{H} = \begin{pmatrix} (\alpha)^{\frac{1}{2}} \mathbf{W}_{\boldsymbol{x}} & \mathbf{0} \\ \mathbf{0} & (1-\alpha)^{\frac{1}{2}} \mathbf{W}_{\boldsymbol{y}} \end{pmatrix} \begin{pmatrix} \mathbf{K}_{\boldsymbol{xx}} & \mathbf{K}_{\boldsymbol{xy}} \\ \mathbf{K}_{\boldsymbol{yx}} & \mathbf{K}_{\boldsymbol{yy}} \end{pmatrix} \begin{pmatrix} (\alpha)^{\frac{1}{2}} \mathbf{W}_{\boldsymbol{x}} & \mathbf{0} \\ \mathbf{0} & (1-\alpha)^{\frac{1}{2}} \mathbf{W}_{\boldsymbol{y}} \end{pmatrix}^T. \tag{65}$$

where

$$\mathbf{W}_{\boldsymbol{x}} = \mathbf{V}_{\boldsymbol{x}}(\mathbf{I}_r - \rho\boldsymbol{\Lambda}_{\boldsymbol{x}}^{-1})^{\frac{1}{2}} \qquad \text{and} \qquad \mathbf{W}_{\boldsymbol{y}} = \mathbf{V}_{\boldsymbol{y}}(\mathbf{I}_r - \rho\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1})^{\frac{1}{2}}. \tag{66}$$

The evaluation of the divergence outside the interval $0 < \alpha < 1$, or when $\beta \neq 1 - \alpha$, is not covered by this formula and, in general, requires knowledge of the eigenvalues of the matrix $\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}}\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}}$. However, different analyses are necessary depending on the dimension of the intersection of the range space of both covariance matrices $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$. In the following, we study the two more general scenarios.

Case (A) The range spaces of $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$ are the same.

In this case $\mathbf{U}_{\boldsymbol{y}}^{\perp}(\mathbf{U}_{\boldsymbol{y}}^{\perp})^T = \mathbf{U}_{\boldsymbol{x}}^{\perp}(\mathbf{U}_{\boldsymbol{x}}^{\perp})^T$, and the eigenvalues of the matrix

$$\begin{aligned} \tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1} &= (\mathbf{C}_{\boldsymbol{x}} + \rho\, \mathbf{U}_{\boldsymbol{x}}^{\perp}(\mathbf{U}_{\boldsymbol{x}}^{\perp})^T)(\mathbf{C}_{\boldsymbol{y}}^{+} + \rho^{-1}\mathbf{U}_{\boldsymbol{x}}^{\perp}(\mathbf{U}_{\boldsymbol{x}}^{\perp})^T) && (67)\\ &= \mathbf{C}_{\boldsymbol{x}}\mathbf{C}_{\boldsymbol{y}}^{+} + \mathbf{U}_{\boldsymbol{x}}^{\perp}(\mathbf{U}_{\boldsymbol{x}}^{\perp})^T && (68) \end{aligned}$$

coincide with the nonzero eigenvalues of $\mathbf{C}_{\boldsymbol{x}}\mathbf{C}_{\boldsymbol{y}}^{+}$ except for $(n - r)$ additional eigenvalues which are equal to 1. Then, using the equivalence between (57) and (60), the divergence reduces to the following form

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) &= L_{AB}^{(\alpha,\beta)}(\mathbf{C}_{\boldsymbol{x}}, \mathbf{C}_{\boldsymbol{y}}) && (69)\\ &= D_{AB}^{(\alpha,\beta)}((\mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{xy}}\mathbf{K}_{\boldsymbol{yy}}^{-1}\mathbf{V}_{\boldsymbol{y}})(\mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{xy}}\mathbf{K}_{\boldsymbol{yy}}^{-1}\mathbf{V}_{\boldsymbol{y}})^T \parallel \mathbf{I}_r). && (70) \end{aligned}$$

Case (B) The range spaces of $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$ are disjoint.

In practice, for $n \gg r$ this is the most probable scenario. In such a case, the r largest eigenvalues of the matrix $\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}$ diverge as ρ tends to zero. Hence, we can not bound above these eigenvalues and, for this reason, it makes no sense to study the case of $\text{sign}(\alpha) \neq \text{sign}(\beta)$, so in this section we assume that $\text{sign}(\alpha) = \text{sign}(\beta)$.

Theorem 3. *When range spaces of $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$ are disjoint and for a sufficiently small value of $\rho > 0$, the AB log-det divergence is closely approximated by the formula*

$$D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}} \,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) \;\approx\; D_{AB}^{(\alpha,\beta)}(\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}^{(\rho)} \parallel \rho\mathbf{I}_r) + D_{AB}^{(\beta,\alpha)}(\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}^{(\rho)} \parallel \rho\mathbf{I}_r), \tag{71}$$

where $\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}^{(\rho)}$ (and respectively $\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}^{(\rho)}$ by interchanging $\boldsymbol{x}$ and $\boldsymbol{y}$) denotes the matrix

$$\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}^{(\rho)} \;=\; \boldsymbol{\Lambda}_{\boldsymbol{x}} - \rho\mathbf{I}_r - \rho^2\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1} - \mathbf{W}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{xy}}\mathbf{W}_{\boldsymbol{y}}\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{W}_{\boldsymbol{y}}^T\mathbf{K}_{\boldsymbol{yx}}\mathbf{W}_{\boldsymbol{x}}. \tag{72}$$

The proof of the theorem is presented in the Appendix G. The eigenvalues of the matrices $\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}^{(\rho)}$ and $\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}^{(\rho)}$, estimate the r largest eigenvalues of $\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}}\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}}$ and of its inverse $\tilde{\mathbf{C}}_{\boldsymbol{x}}^{-\frac{1}{2}}\tilde{\mathbf{C}}_{\boldsymbol{y}}\tilde{\mathbf{C}}_{\boldsymbol{x}}^{-\frac{1}{2}}$, respectively. The relative error in the estimation of these eigenvalues is of order $\mathrm{O}(\rho)$, i.e., it gradually improves as ρ

tend to zero. The approximation is asymptotically exact, and $\mathbf{C}^{(\rho)}_{\boldsymbol{x}|\boldsymbol{y}}$ and $\mathbf{C}^{(\rho)}_{\boldsymbol{x}|\boldsymbol{y}}$ converge respectively to the conditional covariance matrices

$$\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}} \;=\; \lim_{\rho\to 0}\mathbf{C}^{(\rho)}_{\boldsymbol{x}|\boldsymbol{y}} \;=\; \mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}}\mathbf{V}_{\boldsymbol{x}} - (\mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{x}\boldsymbol{y}}\mathbf{V}_{\boldsymbol{y}})(\mathbf{V}_{\boldsymbol{y}}^T\mathbf{K}_{\boldsymbol{y}\boldsymbol{y}}\mathbf{V}_{\boldsymbol{y}})^{-1}(\mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{x}\boldsymbol{y}}\mathbf{V}_{\boldsymbol{y}})^T, \qquad (73)$$

$$\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}} \;=\; \lim_{\rho\to 0}\mathbf{C}^{(\rho)}_{\boldsymbol{y}|\boldsymbol{x}} \;=\; \mathbf{V}_{\boldsymbol{y}}^T\mathbf{K}_{\boldsymbol{y}\boldsymbol{y}}\mathbf{V}_{\boldsymbol{y}} - (\mathbf{V}_{\boldsymbol{y}}^T\mathbf{K}_{\boldsymbol{y}\boldsymbol{x}}\mathbf{V}_{\boldsymbol{x}})(\mathbf{V}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{x}\boldsymbol{x}}\mathbf{V}_{\boldsymbol{x}})^{-1}(\mathbf{V}_{\boldsymbol{y}}^T\mathbf{K}_{\boldsymbol{y}\boldsymbol{x}}\mathbf{V}_{\boldsymbol{x}})^T, \qquad (74)$$

while $\rho\,\mathbf{I}$ converges to the zero matrix.

In the limit, the value of the divergence is not very useful because

$$\lim_{\rho\to 0} D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) = \infty, \qquad (75)$$

though there are some practical ways to circumvent this limitation. For example, when $\alpha = 0$ or $\beta = 0$, the divergence can be scaled by a suitable power of ρ to make it finite (see Section 3.3.1 in [32]). The scaled form of the divergence between the regularized covariance matrices is

$$\mathcal{SD}_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) \;\equiv\; \lim_{\rho\to 0}\rho^{\max\{\alpha,\beta\}} D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}). \qquad (76)$$

Examples of scaled divergences are the following versions of Stein's losses

$$\mathcal{SD}_{AB}^{(0,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) \;=\; \lim_{\rho\to 0}\rho^{\beta} D_{AB}^{(0,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) = \frac{1}{\beta^2}\operatorname{tr}\left((\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}})^{\beta}\right) \ge 0, \qquad \beta > 0, \qquad (77)$$

$$\mathcal{SD}_{AB}^{(\alpha,0)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) \;=\; \lim_{\rho\to 0}\rho^{\alpha} D_{AB}^{(\alpha,0)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) = \frac{1}{\alpha^2}\operatorname{tr}\left((\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}})^{\alpha}\right) \ge 0, \qquad \alpha > 0, \qquad (78)$$

as well as the Jeffrey's KL family of symmetric divergences (*cf.* Equation (23) in [33])

$$\mathcal{SD}_{ABS1}^{(\alpha,0)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) = \lim_{\rho\to 0}\rho^{\alpha} D_{ABS1}^{(\alpha,0)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) = \frac{1}{2\alpha^2}\left(\operatorname{tr}((\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}})^{\alpha}) + \operatorname{tr}((\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}})^{\alpha})\right), \; \alpha > 0. \qquad (79)$$

In other cases, when the scaling is not sufficient to obtain a finite and practical dissimilarity measure, an affine transformation may be used. The idea is to identify the divergent part of $D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}})$ as $\rho \to 0$ and use its value as a reference for the evaluation the dissimilarity. For $\alpha, \beta \ge 0$, the *relative AB log-det dissimilarity measure* is the limiting value of the affine transformation

$$\mathcal{RD}_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) \equiv \lim_{\rho\to 0}\min\{\alpha,\beta\}\left(D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) - \frac{r}{\alpha\beta}\log\frac{\alpha\beta\rho^{-(\alpha+\beta)}}{(\alpha+\beta)^2}\right), \quad \alpha,\beta > 0. \qquad (80)$$

After its extension by continuity (including as special cases $\alpha = 0$ or $\beta = 0$), the function

$$\mathcal{RD}_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) \;=\; \begin{cases} \log\det(\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}) + \dfrac{\alpha}{\beta}\log\det(\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}) & \beta > \alpha \ge 0 \\[2ex] \log\det(\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}) + \log\det(\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}) & \alpha = \beta \ge 0 \\[2ex] \log\det(\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}) + \dfrac{\beta}{\alpha}\log\det(\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}) & \alpha > \beta \ge 0 \end{cases} \qquad (81)$$

provides simple formulas to measure the relative dissimilarity between symmetric positive semidefinite matrices $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$. However, it should be taken into account that, as a consequence of its relative character, this function is not bounded below and can achieve negative values.

7. Modifications and Generalizations of AB Log-Det Divergences and Gamma Matrix Divergences

The divergence (3) discussed in the previous sections can be extended and modified in several ways. It is interesting to note that the positive eigenvalues of $\mathbf{PQ}^{-1}$ play a similar role as the ratios (p_i/q_i) and (q_i/p_i) when used in the wide class of standard discrete divergences, see for example, [11,12]; hence, we can apply such divergences to formulate a modified log-det divergence as a function of the eigenvalues λ_i.

For example, consider the Itakura-Saito distance defined by

$$D_{IS}(\boldsymbol{p}\,\|\,\boldsymbol{q}) \;=\; \sum_i \left(\frac{p_i}{q_i} + \log\frac{q_i}{p_i} - 1\right). \tag{82}$$

It is worth noting that we can generate the large class of divergences or cost functions using Csiszár f-functions [13,24,25]. By replacing p_i/q_i with λ_i and q_i/p_i with λ_i^{-1}, we obtain the log-det divergence for SPD matrices:

$$D_{IS}(\mathbf{P}\,\|\,\mathbf{Q}) \;=\; \sum_{i=1}^{n}(\lambda_i - \log(\lambda_i)) - n, \tag{83}$$

which is consistent with (24) and (26).

As another example, consider the discrete Gamma divergence [11,12] defined by

$$\begin{aligned} D_{AC}^{(\alpha,\beta)}(\boldsymbol{p}\|\boldsymbol{q}) \;&=\; \frac{1}{\beta(\alpha+\beta)}\log\left(\sum_i p_i^{\alpha+\beta}\right) + \frac{1}{\alpha(\alpha+\beta)}\log\left(\sum_i q_i^{\alpha+\beta}\right) - \frac{1}{\alpha\beta}\ln\left(\sum_i p_i^{\alpha} q_i^{\beta}\right) \\ &=\; \frac{1}{\alpha\beta(\alpha+\beta)}\log\frac{\left(\sum_i p_i^{\alpha+\beta}\right)^{\alpha}\left(\sum_i q_i^{\alpha+\beta}\right)^{\beta}}{\left(\sum_i p_i^{\alpha} q_i^{\beta}\right)^{\alpha+\beta}}, \\ &\text{for}\quad \alpha\neq 0,\ \beta\neq 0,\ \alpha+\beta\neq 0, \end{aligned} \tag{84}$$

which when $\alpha = 1$ and $\beta \to -1$, simplifies to the following form [11]:

$$\lim_{\beta\to-1} D_{AC}^{(1,\beta)}(\boldsymbol{p}\,\|\,\boldsymbol{q}) = \frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{q_i}{p_i}\right) + \log\left(\sum_{i=1}^{n}\frac{p_i}{q_i}\right) - \log(n) = \log\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\frac{p_i}{q_i}}{\left(\displaystyle\prod_{i=1}^{n}\frac{p_i}{q_i}\right)^{1/n}}. \tag{85}$$

Hence, by substituting p_i/q_i with λ_i, we derive a new Gamma matrix divergence for SPD matrices:

$$\begin{aligned} D_{CCA}^{(1,0)}(\mathbf{P}\,\|\,\mathbf{Q}) \;&=\; D_{AC}^{(1,-1)}(\mathbf{P}\,\|\,\mathbf{Q}) = \frac{1}{n}\sum_{i=1}^{n}\left(\log\lambda_i^{-1}\right) + \log\left(\sum_{i=1}^{n}\lambda_i\right) - \log(n) \\ &=\; \log\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\lambda_i}{\left(\displaystyle\prod_{i=1}^{n}\lambda_i\right)^{1/n}} = \log\frac{M_1\{\lambda_i\}}{M_0\{\,\lambda_i\}}, \end{aligned} \tag{86}$$

where M_1 denotes the arithmetic mean, and M_0 denotes the geometric mean.

Interestingly, (86) can be expressed equivalently as

$$D_{CCA}^{(1,0)}(\mathbf{P} \,\|\, \mathbf{Q}) = \log(\mathrm{tr}(\mathbf{P}\mathbf{Q}^{-1})) - \frac{1}{n}\log\det(\mathbf{P}\mathbf{Q}^{-1}) - \log(n). \tag{87}$$

Similarly, using the symmetric Gamma divergence defined in [11,12],

$$\begin{aligned} D_{ACS}^{(\alpha,\beta)}(\boldsymbol{p}\|\boldsymbol{q}) &= \frac{1}{\alpha\beta}\log\frac{\left(\sum_i p_i^{\alpha+\beta}\right)\left(\sum_i q_i^{\alpha+\beta}\right)}{\left(\sum_i p_i^{\alpha} q_i^{\beta}\right)\left(\sum_i p_i^{\beta} q_i^{\alpha}\right)}, \\ &\quad \text{for} \quad \alpha \neq 0,\ \beta \neq 0,\ \alpha+\beta \neq 0, \end{aligned} \tag{88}$$

for $\alpha = 1$ and $\beta \to -1$ and by substituting the ratios p_i/q_i with λ_i, we obtain a new Gamma matrix divergence as follows:

$$\begin{aligned} D_{ACS}^{(1,-1)}(\mathbf{P} \,\|\, \mathbf{Q}) &= \log\left((\sum_{i=1}^{n}\lambda_i)(\sum_{i=1}^{n}\lambda_i^{-1})\right) - \log(n)^2 \\ &= \log\left((\frac{1}{n}\sum_{i=1}^{n}\lambda_i)(\frac{1}{n}\sum_{i=1}^{n}\lambda_i^{-1})\right) \\ &= \log\left(M_1\{\lambda_i\}\; M_1\{\lambda_i^{-1}\}\right) \qquad (89) \\ &= \log\frac{M_1\{\lambda_i\}}{M_{-1}\{\lambda_i\}}, \qquad (90) \end{aligned}$$

where $M_{-1}\{\lambda_i\}$ denotes the harmonic mean.

Note that for $n \to \infty$, this formulated divergence can be expressed compactly as

$$D_{ACS}^{(1,-1)}(\mathbf{P} \,\|\, \mathbf{Q}) = \log(E\{\boldsymbol{u}\}\; E\{\boldsymbol{u}^{-1}\}), \tag{91}$$

where $u_i = \{\lambda_i\}$ and $u_i^{-1} = \{\lambda_i^{-1}\}$.

The basic means are defined as follows:

$$M_\gamma(\boldsymbol{\lambda}) = \begin{cases} M_{-\infty} = \min\{\lambda_1,\ldots,\lambda_n\}, & \gamma \to -\infty, \\ M_{-1} = n\left(\sum_{i=1}^{n}\frac{1}{\lambda_i}\right)^{-1}, & \gamma = -1, \\ M_0 = \left(\prod_{i=1}^{n}\lambda_i\right)^{1/n}, & \gamma = 0, \\ M_1 = \frac{1}{n}\sum_{i=1}^{n}\lambda_i, & \gamma = 1, \\ M_2 = \left(\frac{1}{n}\sum_{i=1}^{n}\lambda_i^2\right)^{1/2}, & \gamma = 2, \\ M_\infty = \max\{\lambda_1,\ldots,\lambda_n\}, & \gamma \to \infty, \end{cases} \tag{92}$$

with

$$M_{-\infty} \le M_{-1} \le M_0 \le M_1 \le M_2 \le M_\infty, \tag{93}$$

where equality holds only if all λ_i are equal. By increasing the values of γ, more emphasis is put on large relative errors, i.e., on λ_i whose values are far from one. Depending on the value of γ, we obtain the minimum entry of the vector $\boldsymbol{\lambda}$ (for $\gamma \to -\infty$), its harmonic mean ($\gamma = -1$), the geometric mean ($\gamma = 0$), the arithmetic mean ($\gamma = 1$), the quadratic mean ($\gamma = 2$), and the maximum entry of the vector ($\gamma \to \infty$).

Exploiting the above inequalities for the means, the divergences in (86) and (90) can be heuristically generalized (defined) as follows:

$$D_{CCA}^{(\gamma_2,\gamma_1)}(\mathbf{P} \,\|\, \mathbf{Q}) = \log \frac{M_{\gamma_2}\{\lambda_i\}}{M_{\gamma_1}\{\lambda_i\}}, \tag{94}$$

for $\gamma_2 > \gamma_1$.

The new divergence in (94) is quite general and flexible, and in extreme cases, it takes the following form:

$$D_{CCA}^{(\infty,-\infty)}(\mathbf{P} \,\|\, \mathbf{Q}) = d_H(\mathbf{P} \,\|\, \mathbf{Q}) = \log \frac{M_{\infty}\{\lambda_i\}}{M_{-\infty}\{\lambda_i\}} = \log \frac{\lambda_{max}}{\lambda_{min}}, \tag{95}$$

which is, in fact, a well-known Hilbert projective metric [6,26].

The Hilbert projective metric is extremely simple and suitable for big data because it requires only two (minimum and maximum) eigenvalue computations of the matrix $\mathbf{P}\mathbf{Q}^{-1}$.

The Hilbert projective metric satisfies the following important properties [6,27]:

1. Nonnegativity, $d_H(\mathbf{P} \,\|\, \mathbf{Q}) \geq 0$, and definiteness, $d_H(\mathbf{P} \,\|\, \mathbf{Q}) = 0$, if and only if there exists a $c > 0$ such that $\mathbf{Q} = c\mathbf{P}$.
2. Invariance to scaling:
$$d_H(c_1\mathbf{P} \,\|\, c_2\mathbf{Q}) = d_H(\mathbf{P} \,\|\, \mathbf{Q}), \tag{96}$$
for any $c_1, c_2 > 0$.
3. Symmetry:
$$d_H(\mathbf{P} \,\|\, \mathbf{Q}) = d_H(\mathbf{Q} \,\|\, \mathbf{P})\,. \tag{97}$$
4. Invariance under inversion:
$$d_H(\mathbf{P} \,\|\, \mathbf{Q}) = d_H(\mathbf{P}^{-1} \,\|\, \mathbf{Q}^{-1})\,. \tag{98}$$
5. Invariance under congruence transformations:
$$d_H(\mathbf{A}\mathbf{P}\mathbf{A}^T \,\|\, \mathbf{A}\mathbf{Q}\mathbf{A}^T) = d_H(\mathbf{P} \,\|\, \mathbf{Q}), \tag{99}$$
for any invertible matrix $\mathbf{A}$.
6. Invariance under geodesic (Riemannian) transformations (by taking $\mathbf{A} = \mathbf{P}^{-1/2}$ in (99)):
$$d_H(\mathbf{I} \,\|\, \mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2}) = d_H(\mathbf{P} \,\|\, \mathbf{Q})\,. \tag{100}$$
7. Separability of divergence for the Kronecker product of SPD matrices:
$$d_H(\mathbf{P}_1 \otimes \mathbf{P}_2 \,\|\, \mathbf{Q}_1 \otimes \mathbf{Q}_2) = d_H(\mathbf{P}_1 \,\|\, \mathbf{Q}_1) + d_H(\mathbf{P}_2 \,\|\, \mathbf{Q}_2)\,. \tag{101}$$

8. Scaling of power of SPD matrices:

$$d_H(\mathbf{P}^{\,\omega} \,\|\, \mathbf{Q}^{\omega}) \;=\; |\omega|\; d_H(\mathbf{P} \,\|\, \mathbf{Q}), \tag{102}$$

for any $\omega \neq 0$.
Hence, for $0 < |\omega_1| \leq 1 \leq |\omega_2|$ we have

$$d_H(\mathbf{P}^{\,\omega_1} \,\|\, \mathbf{Q}^{\omega_1}) \leq d_H(\mathbf{P} \,\|\, \mathbf{Q}) \leq d_H(\mathbf{P}^{\,\omega_2} \,\|\, \mathbf{Q}^{\omega_2})\,. \tag{103}$$

9. Scaling under the weighted geometric mean:

$$d_H(\mathbf{P}\#_s\mathbf{Q} \,\|\, \mathbf{P}\#_u\mathbf{Q}) = |s-u|\; d_H(\mathbf{P} \,\|\, \mathbf{Q}), \tag{104}$$

for any $u, s \neq 0$, where

$$\mathbf{P}\#_u\mathbf{Q} = \mathbf{P}^{1/2}(\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,u}\;\mathbf{P}^{1/2}\,. \tag{105}$$

10. Triangular inequality: $d_H(\mathbf{P} \,\|\, \mathbf{Q}) \leq d_H(\mathbf{P} \,\|\, \mathbf{Z}) + d_H(\mathbf{Z} \,\|\, \mathbf{Q})\,.$

These properties can easily be derived and verified. For example, property (9) can easily be derived as follows [6,27]:

$$\begin{aligned} d_H(\mathbf{P}\#_s\mathbf{Q} \,\|\, \mathbf{P}\#_u\mathbf{Q}) &= d_H(\mathbf{P}^{1/2}(\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,s}\;\mathbf{P}^{1/2} \,\|\, (\mathbf{P}^{1/2}(\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,u}\;\mathbf{P}^{1/2}) \\ &= d_H((\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,s} \,\|\, (\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,u}) \\ &= d_H((\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,(s-u)} \,\|\, \mathbf{I}) \\ &= |s-u|\; d_H(\mathbf{P} \,\|\, \mathbf{Q})\,. \end{aligned} \tag{106}$$

In Table 2, we summarize and compare some fundamental properties of three important metric distances: the Hilbert projective metric, Riemannian metric, and LogDet Zero (Bhattacharyya) distance. Since some of these properties are new, we refer to [6,27,28].

7.1. The AB Log-Det Divergence for Noisy and Ill-Conditioned Covariance Matrices

In real-world signal processing and machine learning applications, the SPD sampled matrices can be strongly corrupted by noise and extremely ill conditioned. In such cases, the eigenvalues of the generalized eigenvalue (GEVD) problem $\mathbf{P}\boldsymbol{v}_i = \lambda_i \mathbf{Q}\boldsymbol{v}_i$ can be divided into a signal subspace and noise subspace. The signal subspace is usually represented by the largest eigenvalues (and corresponding eigenvectors), and the noise subspace is usually represented by the smallest eigenvalues (and corresponding eigenvectors), which should be rejected; in other words, in the evaluation of log-det divergences, only the eigenvalues that represent the signal subspace should be taken into account. The simplest approach is to find the truncated dominant eigenvalues by applying the suitable threshold $\tau > 0$; equivalently, find an index $r \leq n$ for which $\lambda_{r+1} \leq \tau$ and perform a summation. For example, truncation reduces the summation in (8) from 1 to r (instead of 1 to n) [22]. The threshold parameter τ can be selected via cross validation.

Table 2. Comparison of the fundamental properties of three basic metric distances: the Riemannian (geodesic) metric (19), LogDet Zero (Bhattacharyya) divergence (21), and the Hilbert projective metric (95). Matrices $\mathbf{P}, \mathbf{Q}, \mathbf{P}_1, \mathbf{P}_2, \mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Z} \in \mathbb{R}^{n\times n}$ are SPD matrices, $\mathbf{A} \in \mathbb{R}^{n\times n}$ is nonsingular, and the matrix $\mathbf{X} \in \mathbb{R}^{n\times r}$ with $r < n$ is full column rank. The scalars satisfy the following conditions: $c > 0$, $c_1, c_2 > 0$; $0 < \omega \leq 1$, $s, u \neq 0$, $\psi = |s - u|$. The geometric means are defined by $\mathbf{P}\#_u\mathbf{Q} = \mathbf{P}^{1/2}(\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,u}\ \mathbf{P}^{1/2}$ and $\mathbf{P}\#\mathbf{Q} = \mathbf{P}\#_{1/2}\mathbf{Q} = \mathbf{P}^{1/2}(\mathbf{P}^{-1/2}\mathbf{Q}\mathbf{P}^{-1/2})^{\,1/2}\ \mathbf{P}^{1/2}$. The Hadamard product of $\mathbf{P}$ and $\mathbf{Q}$ is denoted by $\mathbf{P} \circ \mathbf{Q}$ (*cf.* with [6]).

Riemannian (geodesic) metric	LogDet Zero (Bhattacharyya) div.	Hilbert projective metric
$d_R(\mathbf{P}\\|\mathbf{Q}) = \\|\log(\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})\\|_F$	$d_{\text{Bh}}(\mathbf{P}\\|\mathbf{Q}) = 2\sqrt{\log \dfrac{\det \frac{1}{2}(\mathbf{P}+\mathbf{Q})}{\sqrt{\det(\mathbf{P})\det(\mathbf{Q})}}}$	$d_H(\mathbf{P} \\| \mathbf{Q}) = \log \dfrac{\lambda_{max}\{\mathbf{P}\mathbf{Q}^{-1}\}}{\lambda_{min}\{\mathbf{P}\mathbf{Q}^{-1}\}}$
$d_R(\mathbf{P} \\| \mathbf{Q}) = d_R(\mathbf{Q} \\| \mathbf{P})$	$d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q}) = d_{\text{Bh}}(\mathbf{Q} \\| \mathbf{P})$	$d_H(\mathbf{P} \\| \mathbf{Q}) = d_H(\mathbf{Q} \\| \mathbf{P})$
$d_R(c\mathbf{P} \\| c\mathbf{Q}) = d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(c\mathbf{P} \\| c\mathbf{Q}) = d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(c_1\mathbf{P} \\| c_2\mathbf{Q}) = d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{A}\mathbf{P}\mathbf{A}^T \\| \mathbf{A}\mathbf{Q}\mathbf{A}^T) = d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(\mathbf{A}\mathbf{P}\mathbf{A}^T \\| \mathbf{A}\mathbf{Q}\mathbf{A}^T) = d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{A}\mathbf{P}\mathbf{A}^T \\| \mathbf{A}\mathbf{Q}\mathbf{A}^T) = d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{P}^{-1} \\| \mathbf{Q}^{-1}) = d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(\mathbf{P}^{-1} \\| \mathbf{Q}^{-1}) = d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{P}^{-1} \\| \mathbf{Q}^{-1}) = d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{P}^\omega \\| \mathbf{Q}^\omega) \leq \omega\, d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(\mathbf{P}^\omega \\| \mathbf{Q}^\omega) \leq \sqrt{\omega}\, d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{P}^\omega \\| \mathbf{Q}^\omega) \leq \omega\, d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{P} \\| \mathbf{P}\#_\omega\mathbf{Q}) = \omega\, d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(\mathbf{P} \\| \mathbf{P}\#_\omega\mathbf{Q}) \leq \sqrt{\omega}\, d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{P} \\| \mathbf{P}\#_\omega\mathbf{Q}) = \omega\, d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{Z}\#_\omega\mathbf{P} \\| \mathbf{Z}\#_\omega\mathbf{Q}) \leq \omega\, d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(\mathbf{Z}\#_\omega\mathbf{P} \\| \mathbf{Z}\#_\omega\mathbf{Q}) \leq \sqrt{\omega}\, d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{Z}\#_\omega\mathbf{P} \\| \mathbf{Z}\#_\omega\mathbf{Q}) \leq \omega\, d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{P}\#_s\mathbf{Q} \\| \mathbf{P}\#_u\mathbf{Q}) = \psi\, d_R(\mathbf{P} \\| \mathbf{Q}))$	$d_{\text{Bh}}(\mathbf{P}\#_s\mathbf{Q} \\| \mathbf{P}\#_u\mathbf{Q}) \leq \sqrt{\psi}\, d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{P}\#_s\mathbf{Q} \\| \mathbf{P}\#_u\mathbf{Q}) = \psi\, d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{P} \\| \mathbf{P}\#\mathbf{Q}) = d_R(\mathbf{Q} \\| \mathbf{P}\#\mathbf{Q})$	$d_{\text{Bh}}(\mathbf{P} \\| \mathbf{P}\#\mathbf{Q}) = d_{\text{Bh}}(\mathbf{Q} \\| \mathbf{P}\#\mathbf{Q})$	$d_H(\mathbf{P} \\| \mathbf{P}\#\mathbf{Q}) = d_H(\mathbf{Q} \\| \mathbf{P}\#\mathbf{Q})$
$d_R(\mathbf{X}^T\mathbf{P}\mathbf{X} \\| \mathbf{X}^T\mathbf{Q}\mathbf{X}) \leq d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(\mathbf{X}^T\mathbf{P}\mathbf{X} \\| \mathbf{X}^T\mathbf{Q}\mathbf{X}) \leq d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{X}^T\mathbf{P}\mathbf{X} \\| \mathbf{X}^T\mathbf{Q}\mathbf{X}) \leq d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R(\mathbf{Z} \otimes \mathbf{P} \\| \mathbf{Z} \otimes \mathbf{Q}) = \sqrt{n}\, d_R(\mathbf{P} \\| \mathbf{Q})$	$d_{\text{Bh}}(\mathbf{Z} \otimes \mathbf{P} \\| \mathbf{Z} \otimes \mathbf{Q}) = \sqrt{n}\, d_{\text{Bh}}(\mathbf{P} \\| \mathbf{Q})$	$d_H(\mathbf{Z} \otimes \mathbf{P} \\| \mathbf{Z} \otimes \mathbf{Q}) = d_H(\mathbf{P} \\| \mathbf{Q})$
$d_R^2(\mathbf{P}_1 \otimes \mathbf{P}_2 \\| \mathbf{Q}_1 \otimes \mathbf{Q}_2) =$	$d_{\text{Bh}}(\mathbf{P}_1 \otimes \mathbf{P}_2 \\| \mathbf{Q}_1 \otimes \mathbf{Q}_2)$	$d_H(\mathbf{P}_1 \otimes \mathbf{P}_2 \\| \mathbf{Q}_1 \otimes \mathbf{Q}_2)$
$= n\, d_R^2(\mathbf{P}_1 \\| \mathbf{Q}_1) + n\, d_R^2(\mathbf{P}_2 \\| \mathbf{Q}_2) +$	$\geq d_{\text{Bh}}(\mathbf{P}_1 \circ \mathbf{P}_2 \\| \mathbf{Q}_1 \circ \mathbf{Q}_2)$	$= d_H(\mathbf{P}_1 \\| \mathbf{Q}_1) + d_H(\mathbf{P}_2 \\| \mathbf{Q}_2)$
$2\ \log\det(\mathbf{P}_1\mathbf{Q}_1^{-1})\ \log\det(\mathbf{P}_2\mathbf{Q}_2^{-1})$		

Recent studies suggest that the real signal subspace covariance matrices can be better represented by truncating the eigenvalues. A popular and relatively simple method applies a thresholding and shrinkage rule to the eigenvalues [35]:

$$\widetilde{\lambda}_i = \lambda_i \max\{(1 - \frac{\tau^\gamma}{\lambda^\gamma}), 0\}, \tag{107}$$

where any eigenvalue smaller than the specific threshold is set to zero, and the remaining eigenvalues are shrunk. Note that the smallest eigenvalues are shrunk more than the largest one. For $\gamma = 1$, we obtain a standard soft thresholding, and for $\gamma \to \infty$ a standard hard thresholding is obtained [36]. The optimal threshold $\tau > 0$ can be estimated along with the parameter $\gamma > 0$ using cross validation. However, a more practical and efficient method is to apply the Generalized Stein Unbiased Risk Estimate (GSURE) method even if the variance of the noise is unknown (for more details, we refer to [35] and the references therein).

In this paper, we propose an alternative approach in which the bias generated by noise is reduced by suitable choices of α and β [12]. Instead of using the eigenvalues λ_i of $\mathbf{PQ}^{-1}$ or its inverse, we use regularized or shrinked eigenvalues [35–37]. For example, in light of (8), we can use the following shrinked eigenvalues:

$$\widetilde{\lambda}_i = \left(\frac{\alpha \lambda_i^\beta + \beta \lambda_i^{-\alpha}}{\alpha + \beta} \right)^{\frac{1}{\alpha\beta}} \geq 1, \quad \text{for} \;\; \alpha, \beta \neq 0, \;\; \alpha, \beta > 0 \;\; \text{or} \;\; \alpha, \beta < 0, \tag{108}$$

which play a similar role as the ratios (p_i/q_i) $(p_i \geq q_i)$, which are used in the standard discrete divergences [11,12]. It should be noted that equalities $\widetilde{\lambda}_i = 1, \quad \forall i$ hold only if all λ_i of $\mathbf{PQ}^{-1}$ are equal to one, which occurs only if $\mathbf{P} = \mathbf{Q}$. For example, the new Gamma divergence in (94) can be formulated even more generally as

$$D_{CCA}^{(\gamma_2,\gamma_1)}(\mathbf{P} \,||\, \mathbf{Q}) = \log \frac{M_{\gamma_2}\{\widetilde{\lambda}_i\}}{M_{\gamma_1}\{\widetilde{\lambda}_i\}}, \tag{109}$$

where $\gamma_2 > \gamma_1$, and $\widetilde{\lambda}_i$ are the regularized or optimally shrinked eigenvalues.

8. Divergences of Multivariate Gaussian Densities and Differential Relative Entropies of Multivariate Normal Distributions

In this section, we show the links or relationships between a family of continuous Gamma divergences and AB log-det divergences for multivariate Gaussian densities.

Consider the two multivariate Gaussian (normal) distributions:

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{P}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \mathbf{P}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) \right), \tag{110}$$

$$q(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n \det \mathbf{Q}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \mathbf{Q}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \right), \;\; \boldsymbol{x} \in \mathbb{R}^n, \tag{111}$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^n$ and $\boldsymbol{\mu}_2 \in \mathbb{R}^n$ are mean vectors, and $\mathbf{P} = \boldsymbol{\Sigma}_1 \in \mathbb{R}^{n \times n}$ and $\mathbf{Q} = \boldsymbol{\Sigma}_2 \in \mathbb{R}^{n \times n}$ are the covariance matrices of $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, respectively.

Furthermore, consider the Gamma divergence for these distributions:

$$D_{AC}^{(\alpha,\beta)}(p(\boldsymbol{x})\|q(\boldsymbol{x})) = \frac{1}{\alpha\beta(\alpha+\beta)} \log \frac{\left(\int_\Omega p^{\alpha+\beta}(\boldsymbol{x})\, d\boldsymbol{x}\right)^\alpha \left(\int_\Omega q^{\alpha+\beta}(\boldsymbol{x})\, d\boldsymbol{x}\right)^\beta}{\left(\int_\Omega p^\alpha(\boldsymbol{x})\, q^\beta(\boldsymbol{x})\, d\boldsymbol{x}\right)^{\alpha+\beta}} \quad \text{for} \quad \alpha \neq 0,\ \beta \neq 0,\ \alpha+\beta \neq 0, \tag{112}$$

which generalizes a family of Gamma divergences [11,12].

Theorem 4. *The Gamma divergence in (112) for multivariate Gaussian densities (110) and (111) can be expressed in closed form as follows:*

$$\begin{aligned} D_{AC}^{(\alpha,\beta)}(p(\boldsymbol{x})\|q(\boldsymbol{x})) &= \frac{1}{2} D_{AB}^{(\beta,\alpha)}\left((\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})^{\frac{1}{\alpha+\beta}} \| \mathbf{I}\right) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T (\alpha\mathbf{Q} + \beta\mathbf{P})^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \\ &= \frac{1}{2\alpha\beta} \log \frac{\det\left(\frac{\alpha}{\alpha+\beta}\mathbf{Q} + \frac{\beta}{\alpha+\beta}\mathbf{P}\right)}{\det(\mathbf{Q})^{\frac{\alpha}{\alpha+\beta}} \det(\mathbf{P})^{\frac{\beta}{\alpha+\beta}}} \\ &\quad + \frac{1}{2(\alpha+\beta)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left(\frac{\alpha}{\alpha+\beta}\mathbf{Q} + \frac{\beta}{\alpha+\beta}\mathbf{P}\right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \end{aligned} \tag{113}$$

for $\alpha > 0$ *and* $\beta > 0$.

The proof is provided in Appendix H. Note that for $\alpha + \beta = 1$, the first term in the right-hand-side of (113) also simplifies as

$$\frac{1}{2} D_{AB}^{(\beta,\alpha)}\left((\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})^{\frac{1}{\alpha+\beta}} \| \mathbf{I}\right)\Big|_{\beta=1-\alpha} = \frac{1}{2} D_{AB}^{(1-\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{2} D_{A}^{(1-\alpha)}(\mathbf{P}\|\mathbf{Q}). \tag{114}$$

Observe that Formula (113) consists of two terms: the first term is expressed via the AB log-det divergence, which measures the similarity between two covariance or precision matrices and is independent from the mean vectors, while the second term is a quadratic form expressed by the Mahalanobis distance, which represents the distance between the means (weighted by the covariance matrices) of multivariate Gaussian distributions. Note that the second term is zero when the mean values $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ coincide.

Theorem 4 is a generalization of the following well-known results:

1. For $\alpha = 1$ and $\beta = 0$ and as $\beta \to 0$, the Kullback-Leibler divergence can be expressed as [5,38]

$$\begin{aligned} \lim_{\beta \to 0} D_{AC}^{(1,\beta)}(p(\boldsymbol{x})\,\|\,q(\boldsymbol{x})) &= D_{KL}(p(\boldsymbol{x})\|q(\boldsymbol{x})) = \int_\Omega p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x} \\ &= \frac{1}{2}\left(\operatorname{tr}(\mathbf{P}\mathbf{Q}^{-1}) - \log\det(\mathbf{P}\mathbf{Q}^{-1}) - n\right) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{Q}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \end{aligned} \tag{115}$$

   where the last term represents the Mahalanobis distance, which becomes zero for zero-mean distributions $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = 0$.

2. For $\alpha = \beta = 0.5$ we have the Bhattacharyya distance [5,39]

$$D_{AC}^{(0.5,0.5)}(p(\boldsymbol{x})\|q(\boldsymbol{x})) = \frac{1}{2} d_{\rm Bh}^2(p(\boldsymbol{x})\|q(\boldsymbol{x})) = -4 \log \int_\Omega \sqrt{p(\boldsymbol{x})q(\boldsymbol{x})} d\boldsymbol{x} \tag{116}$$

$$= 2 \log \frac{\det \dfrac{\mathbf{P}+\mathbf{Q}}{2}}{\sqrt{\det \mathbf{P} \det \mathbf{Q}}} + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left[\frac{\mathbf{P}+\mathbf{Q}}{2}\right]^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

3. For $\alpha + \beta = 1$ and $0 < \alpha < 1$, the closed form expression for the Rényi divergence is obtained [5,32,40]:

$$D_A(p\|q) = -\frac{1}{\alpha(1-\alpha)} \log \int_\Omega p^{\,\alpha}(\boldsymbol{x})\, q^{\,1-\alpha}(\boldsymbol{x}) d\boldsymbol{x} \tag{117}$$

$$= \frac{1}{2\alpha(1-\alpha)} \log \frac{\det(\alpha\mathbf{Q} + (1-\alpha)\mathbf{P})}{\det(\mathbf{Q}^\alpha\, \mathbf{P}^{1-\alpha})} + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left[\alpha\mathbf{Q} + (1-\alpha)\mathbf{P}\right]^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

4. For $\alpha = \beta = 1$, the Gamma-divergences reduce to the Cauchy-Schwartz divergence:

$$D_{CS}(p(\boldsymbol{x}) \,\|\, q(\boldsymbol{x})) = -\log \frac{\displaystyle\int p(\boldsymbol{x})\; q(\boldsymbol{x})\; d\mu(\boldsymbol{x})}{\left(\displaystyle\int p^2(\boldsymbol{x}) d\mu(\boldsymbol{x})\right)^{1/2} \left(\displaystyle\int q^2(\boldsymbol{x}) d\mu(\boldsymbol{x})\right)^{1/2}} \tag{118}$$

$$= \frac{1}{2} \log \frac{\det \dfrac{\mathbf{P}+\mathbf{Q}}{2}}{\sqrt{\det \mathbf{Q} \det \mathbf{P}}} + \frac{1}{4}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \left(\frac{\mathbf{P}+\mathbf{Q}}{2}\right)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,.$$

Similar formulas can be derived for the symmetric Gamma divergence for two multivariate Gaussian distributions. Furthermore, analogous expressions can be derived for Elliptical Gamma distributions (EGD) [41], which facilitate more flexible modeling than standard multivariate Gaussian distributions.

8.1. Multiway Divergences for Multivariate Normal Distributions with Separable Covariance Matrices

Recently, there has been growing interest in the analysis of tensors or multiway arrays [42–45]. One of the most important applications of multiway tensor analysis and multilinear distributions, is magnetic resonance imaging (MRI) (we refer to [46] and the references therein). For multiway arrays, we often use multilinear (array or tensor) normal distributions that correspond to the multivariate normal (Gaussian) distributions in (110) and (111) with common means $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and separable (Kronecker structured) covariance matrices:

$$\bar{\mathbf{P}} = \sigma_P^2 \left(\mathbf{P}_1 \otimes \mathbf{P}_2 \otimes \cdots \otimes \mathbf{P}_K\right) \in \mathbb{R}^{N \times N} \tag{119}$$

$$\bar{\mathbf{Q}} = \sigma_Q^2 \left(\mathbf{Q}_1 \otimes \mathbf{Q}_2 \otimes \cdots \otimes \mathbf{Q}_K\right) \in \mathbb{R}^{N \times N}, \tag{120}$$

where $\mathbf{P}_k \in \mathbb{R}^{n_k \times n_k}$ and $\mathbf{Q}_k \in \mathbb{R}^{n_k \times n_k}$ for $k = 1, 2, \ldots, K$ are SPD matrices, usually normalized so that $\det \mathbf{P}_k = \det \mathbf{Q}_k = 1$ for each k and $N = \prod_{k=1}^K n_k$ [45].

One of the main advantages of the separable Kronecker model is the significant reduction in the number of variance-covariance parameters [42]. Usually, such separable covariance matrices are sparse and very large-scale. The challenge is to design an efficient and relatively simple dissimilarity measure for big data between two zero-mean multivariate (or multilinear) normal distributions ((110) and (111)). Because of its unique properties, the Hilbert projective metric is a good candidate; in particular, for separable Kronecker structured covariances, it can be expressed very simply as

$$D_H(\bar{\mathbf{P}} \,\|\, \bar{\mathbf{Q}}) = \sum_{k=1}^{K} D_H(\mathbf{P}_k \,\|\, \mathbf{Q}_k) = \sum_{k=1}^{K} \log \frac{\widetilde{\lambda}_{max}^{(k)}}{\widetilde{\lambda}_{min}^{(k)}} = \log \prod_{k=1}^{K} \left(\frac{\widetilde{\lambda}_{max}^{(k)}}{\widetilde{\lambda}_{min}^{(k)}} \right), \tag{121}$$

where $\widetilde{\lambda}_{max}^{(k)}$ and $\widetilde{\lambda}_{min}^{(k)}$ are the (shrinked) maximum and minimum eigenvalues of the (relatively small) matrices $\mathbf{P}_k \mathbf{Q}_k^{-1}$ for $k = 1, 2, \ldots, K$, respectively. We refer to this divergence as the multiway Hilbert metric. This metric has many attractive properties, especially invariance under multilinear transformations.

Using the fundamental properties of divergence and SPD matrices, we derive other multiway log-det divergences. For example, the multiway Stein's loss can be obtained:

$$\begin{aligned} D_{MSL}(\bar{\mathbf{P}}, \bar{\mathbf{Q}}) &= 2\, D_{KL}(p(\boldsymbol{x}) \,\|\, q(\boldsymbol{x})) = D_{AB}^{(0,1)}(\bar{\mathbf{P}} \,\|\, \bar{\mathbf{Q}}) && (122) \\ &= \operatorname{tr}\left(\bar{\mathbf{P}}\bar{\mathbf{Q}}^{-1}\right) - \log\det(\bar{\mathbf{P}}\bar{\mathbf{Q}}^{-1}) - N \\ &= \frac{\sigma_P^2}{\sigma_Q^2} \left(\prod_{k=1}^{K} \operatorname{tr}(\mathbf{P}_k \mathbf{Q}_k^{-1}) \right) - \sum_{k=1}^{K} \frac{N}{n_k} \log\det(\mathbf{P}_k \mathbf{Q}_k^{-1}) - N \log \left(\frac{\sigma_P^2}{\sigma_Q^2} \right) - N. && (123) \end{aligned}$$

Note that under the constraint that $\det \mathbf{P}_k = \det \mathbf{Q}_k = 1$, this simplifies to

$$\begin{aligned} D_{MSL}(\bar{\mathbf{P}} \,\|\, \bar{\mathbf{Q}}) &= \operatorname{tr}\left(\bar{\mathbf{P}}\bar{\mathbf{Q}}^{-1}\right) - \log\det(\bar{\mathbf{P}}\bar{\mathbf{Q}}^{-1}) - N && (124) \\ &= \frac{\sigma_P^2}{\sigma_Q^2} \left(\prod_{k=1}^{K} \operatorname{tr}(\mathbf{P}_k \mathbf{Q}_k^{-1}) \right) - N \log \left(\frac{\sigma_P^2}{\sigma_Q^2} \right) - N, \end{aligned}$$

which is different from the multiway Stein's loss recently proposed by Gerard and Hoff [45].

Similarly, if $\det \mathbf{P}_k = \det \mathbf{Q}_k = 1$ for each $k = 1, 2, \ldots, K$, we can derive the multiway Riemannian metric as follows:

$$d_R^2(\bar{\mathbf{P}} \,\|\, \bar{\mathbf{Q}}) = N \log^2 \frac{\sigma_P^2}{\sigma_Q^2} + \sum_{k=1}^{K} \frac{N}{n_k}\, d_R^2(\mathbf{P}_k \,\|\, \mathbf{Q}_k)\,. \tag{125}$$

The above multiway divergences are derived using the following properties:

$$\begin{aligned} \bar{\mathbf{P}}\bar{\mathbf{Q}}^{-1} &= (\mathbf{P}_1 \otimes \mathbf{P}_2 \otimes \cdots \otimes \mathbf{P}_K)(\mathbf{Q}_1^{-1} \otimes \mathbf{Q}_2^{-1} \otimes \cdots \otimes \mathbf{Q}_K^{-1}) \\ &= \mathbf{P}_1\mathbf{Q}_1^{-1} \otimes \mathbf{P}_2\mathbf{Q}_2^{-1} \otimes \cdots \otimes \mathbf{P}_K\mathbf{Q}_K^{-1}, && (126) \\ \operatorname{tr}(\bar{\mathbf{P}}\bar{\mathbf{Q}}^{-1}) &= \operatorname{tr}(\mathbf{P}_1\mathbf{Q}_1^{-1} \otimes \mathbf{P}_2\mathbf{Q}_2^{-1} \otimes \cdots \otimes \mathbf{P}_K\mathbf{Q}_K^{-1}) = \prod_{k=1}^{K} \operatorname{tr}(\mathbf{P}_k\mathbf{Q}_k^{-1}), && (127) \\ \det(\bar{\mathbf{P}}\bar{\mathbf{Q}}^{-1}) &= \det(\mathbf{P}_1\mathbf{Q}_1^{-1} \otimes \mathbf{P}_2\mathbf{Q}_2^{-1} \otimes \cdots \otimes \mathbf{P}_K\mathbf{Q}_K^{-1}) = \prod_{k=1}^{K} (\det(\mathbf{P}_k\mathbf{Q}_k^{-1}))^{N/n_k}. && (128) \end{aligned}$$

and the basic property: If the eigenvalues $\{\lambda_i\}$ and $\{\theta_j\}$ are eigenvalues with corresponding eigenvectors $\{\boldsymbol{v}_i\}$ and $\{\boldsymbol{u}_j\}$ for SPD matrices $\mathbf{A}$ and $\mathbf{B}$, respectively, then $\mathbf{A} \otimes \mathbf{B}$ has eigenvalues $\{\lambda_i \theta_j\}$ with corresponding eigenvectors $\{\boldsymbol{v}_i \otimes \boldsymbol{u}_j\}$.

Other possible extensions of the AB and Gamma matrix divergences to separable multiway divergences for multilinear normal distributions under additional constraints and normalization conditions will be discussed in future works.

9. Conclusions

In this paper, we presented novel (dis)similarity measures; in particular, we considered the Alpha-Beta and Gamma log-det divergences (and/or their square-roots) that smoothly connect or unify a wide class of existing divergences for SPD matrices. We derived numerous results that uncovered or unified theoretic properties and qualitative similarities between well-known divergences and new divergences. The scope of the results presented in this paper is vast, especially since the parameterized Alpha-Beta and Gamma log-det divergence functions include several efficient and useful divergences, including those based on relative entropies, the Riemannian metric (AIRM), S-divergence, generalized Jeffreys KL (KLDM), Stein's loss, and Hilbert projective metric. Various links and relationships between divergences were also established. Furthermore, we proposed several multiway log-det divergences for tensor (array) normal distributions.

Acknowledgments

Part of this work was supported by the Spanish Government under MICINN projects TEC2014-53103, TEC2011-23559, and by the Regional Government of Andalusia under Grant TIC-7869.

Author Contributions

First two authors contributed equally to this work. Andrzej Cichocki has coordinated this study and wrote most of the sections 1-3 and 7-8. Sergio Cruces wrote most of the sections 4, 5, and 6. He also provided most of the final rigorous proofs presented in Appendices. Shun-ichi Amari proved the fundamental property (16) that the Riemannian metric is the same for all AB log-det divergences and critically revised the paper by providing inspiring comments. All authors have read and approved the final manuscript.

Appendices

A. Basic operations for positive definite matrices

Functions of positive definite matrices frequently appear in many research areas, for an introduction we refer the reader to Chapter 11 in [31]. Consider a positive definite matrix $\mathbf{P}$ of rank n with eigendecomposition $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. The matrix function $f(\mathbf{P})$ is defined as

$$f(\mathbf{P}) = \mathbf{V} f(\mathbf{\Lambda}) \mathbf{V}^T, \tag{129}$$

where $f(\mathbf{\Lambda}) \equiv \operatorname{diag}\left(f(\lambda_1), \ldots, f(\lambda_n)\right)$. With the help of this definition, the following list of well-known properties can be easily obtained:

$$\begin{aligned}
\log(\det\ \mathbf{P}) &= \operatorname{tr}\log(\mathbf{P}), & (130)\\
(\det\mathbf{P})^{\alpha} &= \det(\mathbf{P}^{\alpha}), & (131)\\
\det(\mathbf{P}^{\alpha}) &= \det(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{T})^{\alpha} = \det(\mathbf{V})\det(\boldsymbol{\Lambda}^{\alpha})\det(\mathbf{V}^{T}) = \prod_{i=1}^{n}\lambda_i^{\alpha}, & (132)\\
\operatorname{tr}(\mathbf{P}^{\alpha}) &= \operatorname{tr}(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^{T})^{\alpha} = \operatorname{tr}(\mathbf{V}\mathbf{V}^{T}\boldsymbol{\Lambda}^{\alpha}) = \sum_{i=1}^{n}\lambda_i^{\alpha}, & (133)\\
\mathbf{P}^{\alpha+\beta} &= \mathbf{P}^{\alpha}\,\mathbf{P}^{\beta}, & (134)\\
(\mathbf{P}^{\,\alpha})^{\,\beta} &= \mathbf{P}^{\,\alpha\,\beta}, & (135)\\
\mathbf{P}^{\,0} &= \mathbf{I}, & (136)\\
(\det\mathbf{P})^{\alpha+\beta} &= \det(\mathbf{P}^{\alpha})\det(\mathbf{P}^{\beta}), & (137)\\
\det((\mathbf{P}\mathbf{Q}^{-1})^{\alpha}) &= [\det(\mathbf{P})\det(\mathbf{Q}^{-1})]^{\alpha} = \det(\mathbf{P}^{\alpha})\det(\mathbf{Q}^{-\alpha}), & (138)\\
\frac{\partial}{\partial\alpha}(\mathbf{P}^{\alpha}) &= \mathbf{P}^{\alpha}\log(\mathbf{P}), & (139)\\
\frac{\partial}{\partial\alpha}\log\left[\det(\mathbf{P}(\alpha))\right] &= \operatorname{tr}\left(\mathbf{P}^{-1}\frac{\partial\mathbf{P}}{\partial\alpha}\right), & (140)\\
\log(\det(\mathbf{P}\otimes\mathbf{Q})) &= n\log(\det\mathbf{P}) + n\log(\det\mathbf{Q}), & (141)\\
\operatorname{tr}(\mathbf{P}) - \log\det(\mathbf{P}) &\geq n. & (142)
\end{aligned}$$

B. Extension of $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ for $(\alpha,\beta)\in\mathbb{R}^2$

Remark 1. *Equation (3) is only well defined in the first and third quadrants of the (α,β)-plane. Outside these regions, where α and β have opposite signs (i.e., $\alpha > 0$ and $\beta < 0$ or $\alpha < 0$ and $\beta > 0$), the divergence can be complex valued.*

This undesirable behavior can be avoided with the help of the truncation operator

$$[x]_+ = \begin{cases} x & x \geq 0, \\ 0, & x < 0, \end{cases} \qquad (143)$$

which prevents the arguments of the logarithms from being negative.
The new definition of the AB log-det divergence is

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\alpha\beta}\log\left[\det\frac{\alpha(\mathbf{P}\mathbf{Q}^{-1})^{\beta} + \beta(\mathbf{P}\mathbf{Q}^{-1})^{-\alpha}}{\alpha+\beta}\right]_+ \qquad (144)$$

$$\text{for } \alpha\neq 0,\quad \beta\neq 0,\quad \alpha+\beta\neq 0,$$

which is compatible with the previous definition in the first and third quadrants of the (α,β)-plane. It is also well defined in the second and fourth quadrants except for the special cases when $\alpha = 0,\ \beta = 0,$

and $\alpha + \beta = 0$, which is where the formula is undefined. By enforcing continuity, we can explicitly define the AB log-det divergence on the entire (α, β)-plane as follows:

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \begin{cases} \dfrac{1}{\alpha\beta} \log\det \left[\dfrac{\alpha(\mathbf{P}\mathbf{Q}^{-1})^{\beta} + \beta(\mathbf{Q}\mathbf{P}^{-1})^{\alpha}}{\alpha+\beta} \right]_{+} & \text{for } \ \alpha, \beta \neq 0, \ \ \alpha+\beta \neq 0, \\[4em] \dfrac{1}{\alpha^2} \left[\operatorname{tr}\left((\mathbf{Q}\mathbf{P}^{-1})^{\alpha} - \mathbf{I} \right) - \alpha \log\det(\mathbf{Q}\mathbf{P}^{-1}) \right] & \text{for } \ \alpha \neq 0, \ \beta = 0, \\[4em] \dfrac{1}{\beta^2} \left[\operatorname{tr}\left((\mathbf{P}\mathbf{Q}^{-1})^{\beta} - \mathbf{I} \right) - \beta \log\det(\mathbf{P}\mathbf{Q}^{-1}) \right] & \text{for } \ \alpha = 0, \ \beta \neq 0, \\[4em] \dfrac{1}{\alpha^2} \log\det[(\mathbf{P}\mathbf{Q}^{-1})^{-\alpha}(\mathbf{I} + \log(\mathbf{P}\mathbf{Q}^{-1})^{\alpha})]_{+}^{-1} & \text{for } \alpha = -\beta, \\[4em] \dfrac{1}{2} \operatorname{tr} \log^2(\mathbf{P}\mathbf{Q}^{-1}) = \dfrac{1}{2} || \log(\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2}) ||_F^2 & \text{for } \ \alpha, \ \beta = 0. \end{cases} \tag{145}$$

C. Eigenvalues Domain for Finite $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$

In this section, we assume that λ_i, an eigenvalue of $\mathbf{P}\mathbf{Q}^{-1}$, satisfies $0 \leq \lambda_i \leq \infty$ for all $i = 1, \ldots, n$. We will determine the bounds of the eigenvalues of $\mathbf{P}\mathbf{Q}^{-1}$ that prevent the AB log-det divergence from being infinite. First, recall that

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \ = \ \frac{1}{\alpha\beta} \sum_{i=1}^{n} \log \left[\frac{\alpha\lambda_i^{\beta} + \beta\lambda_i^{-\alpha}}{\alpha+\beta} \right]_{+}, \ \ \alpha, \ \beta, \ \alpha+\beta \neq 0. \tag{146}$$

We assume that $0 \leq \lambda_i \leq \infty$ for all i. For the divergence to be finite, the arguments of the logarithms in the previous expression must be positive. This happens when

$$\frac{\alpha\lambda_i^{\beta} + \beta\lambda_i^{-\alpha}}{\alpha+\beta} > 0 \qquad \forall i, \tag{147}$$

which is always true when $\alpha, \beta > 0$ or when $\alpha, \beta < 0$. On the contrary, when $\operatorname{sign}(\alpha\beta) = -1$, we have the following two cases. In the first case when $\alpha > 0$, we initially solve for $\lambda_i^{\alpha+\beta}$ and later for λ_i to obtain

$$\frac{\lambda_i^{\alpha+\beta}}{\alpha+\beta} > \frac{-\beta}{\alpha(\alpha+\beta)} = \left|\frac{\beta}{\alpha}\right| \frac{1}{\alpha+\beta} \quad \longrightarrow \quad \lambda_i > \left|\frac{\beta}{\alpha}\right|^{\frac{1}{\alpha+\beta}} \qquad \forall i, \ \text{for } \alpha > 0 \text{ and } \beta < 0. \tag{148}$$

In the second case when $\alpha < 0$, we obtain

$$\frac{\lambda_i^{\alpha+\beta}}{\alpha+\beta} < \frac{-\beta}{\alpha(\alpha+\beta)} = \left|\frac{\beta}{\alpha}\right| \frac{1}{\alpha+\beta} \quad \longrightarrow \quad \lambda_i < \left|\frac{\beta}{\alpha}\right|^{\frac{1}{\alpha+\beta}} \qquad \forall i, \ \text{for } \alpha < 0 \text{ and } \beta > 0. \tag{149}$$

Using $\text{sign}(\alpha\beta) = -1$, we can solve for $\lambda_i^{\alpha+\beta}$, which yields

$$\frac{\lambda_i^{\alpha+\beta}}{\alpha+\beta} \;>\; \left|\frac{\beta}{\alpha}\right| \frac{1}{\alpha+\beta} \qquad \forall i. \tag{150}$$

Solving again for λ_i, we see that

$$\lambda_i \;>\; \left|\frac{\beta}{\alpha}\right|^{\frac{1}{\alpha+\beta}} \qquad \forall i, \text{ for } \alpha > 0 \text{ and } \beta < 0, \tag{151}$$

and

$$\lambda_i \;<\; \left|\frac{\beta}{\alpha}\right|^{\frac{1}{\alpha+\beta}} \qquad \forall i, \text{ for } \alpha < 0 \text{ and } \beta > 0. \tag{152}$$

In the limit, when $\alpha \to -\beta \neq 0$, these bounds simplify to

$$\lim_{\alpha\to-\beta} \left|\frac{\beta}{\alpha}\right|^{\frac{1}{\alpha+\beta}} = e^{-1/\alpha} \qquad \forall i, \text{ for } \beta \neq 0. \tag{153}$$

On the other hand, when $\alpha \to 0$ or when $\beta \to 0$, the bounds disappear. The lower-bounds converge to 0, while the upper-bounds converge to ∞, leading to the trivial inequalities $0 < \lambda_i < \infty$.

This concludes the determination of the domain of the eigenvalues that result in a finite divergence. Outside this domain, we expect $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = \infty$. A complete picture of bounds for different values of α and β is shown in Figure 1.

D. Proof of the Nonnegativity of $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$

The AB log-det divergence is separable; it is the sum of the individual divergences of the eigenvalues from unity, *i.e.*,

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \;=\; \sum_{i=1}^{n} D_{AB}^{(\alpha,\beta)}(\lambda_i\|1), \tag{154}$$

where

$$D_{AB}^{(\alpha,\beta)}(\lambda_i\|1) \;=\; \frac{1}{\alpha\beta} \log\left[\frac{\alpha\lambda_i^{\beta} + \beta\lambda_i^{-\alpha}}{\alpha+\beta}\right]_+, \quad \alpha,\ \beta,\ \alpha+\beta \neq 0. \tag{155}$$

We prove the nonnegativity of $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ by showing that the divergence of each of the eigenvalues $D_{AB}^{(\alpha,\beta)}(\lambda_i\|1)$ is nonnegative and minimal at $\lambda_i = 1$.

First, note that the only critical point of the criterion is obtained when $\lambda_i = 1$. This can be shown by setting the derivative of the criterion equal to zero, *i.e.*,

$$\frac{\partial D_{AB}^{(\alpha,\beta)}(\lambda_i\|1)}{\partial\lambda_i} \;=\; \frac{\lambda_i^{\alpha+\beta} - 1}{\alpha\lambda_i^{\alpha+\beta+1} + \beta\lambda_i} = 0, \tag{156}$$

and solving for λ_i.

Next, we show that the sign of the derivative only changes at the critical point $\lambda_i = 1$. If we rewrite

$$\frac{\partial D_{AB}^{(\alpha,\beta)}(\lambda_i \| 1)}{\partial \lambda_i} \;=\; \left(\frac{\lambda_i^{\alpha+\beta}-1}{\alpha+\beta}\right)\left(\lambda_i \frac{\alpha\lambda_i^{\alpha+\beta}+\beta}{\alpha+\beta}\right)^{-1}, \tag{157}$$

and observe that the condition for the divergence to be finite enforces $\frac{\alpha\lambda_i^{\alpha+\beta}+\beta}{\alpha+\beta} > 0$, then it follows that

$$\operatorname{sign}\left\{\frac{\partial D_{AB}^{(\alpha,\beta)}(\lambda_i \| 1)}{\partial \lambda_i}\right\} \equiv \operatorname{sign}\left\{\frac{\lambda_i^{\alpha+\beta}-1}{\alpha+\beta}\right\} = \begin{cases} -1 & \text{for } \lambda_i < 1, \\ 0, & \text{for } \lambda_i = 1, \\ +1 & \text{for } \lambda_i > 1. \end{cases} \tag{158}$$

Since the derivative is strictly negative for $\lambda_i < 1$ and strictly positive for $\lambda_i > 1$, the critical point at $\lambda_i = 1$ is the global minimum of $D_{AB}^{(\alpha,\beta)}(\lambda_i \| 1)$. From this result, the nonnegativity of the divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \geq 0$ easily follows. Moreover, $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = 0$ only when $\lambda_i = 1$ for $i = 1, \ldots, n$, which concludes the proof of the Theorems 1 and 2.

E. Derivation of the Riemannian Metric

We calculate $D_{AB}^{(\alpha,\beta)}(\mathbf{P} + d\mathbf{P} \,\|\, \mathbf{P})$ using the Taylor expansion when $d\mathbf{P}$ is small, *i.e.*,

$$(\mathbf{P} + d\mathbf{P})\mathbf{P}^{-1} = \mathbf{I} + d\mathbf{Z}, \tag{159}$$

where

$$\begin{aligned} d\mathbf{Z} &= d\mathbf{P}\mathbf{P}^{-1}, \\ \alpha[(\mathbf{P} + d\mathbf{P})\mathbf{P}^{-1}]^{\beta} &= \alpha\mathbf{I} + \alpha\beta\, d\mathbf{Z} + \frac{\alpha\beta(\beta-1)}{2}\, d\mathbf{Z}\, d\mathbf{Z} + O(|d\mathbf{Z}|^3). \end{aligned}$$

Similar calculations hold for $\beta[(\mathbf{P} + d\mathbf{P})\mathbf{P}^{-1}]^{-\alpha}$, and

$$\alpha[(\mathbf{P} + d\mathbf{P})\mathbf{P}^{-1}]^{\beta} + \beta[(\mathbf{P} + d\mathbf{P})\mathbf{P}^{-1}]^{-\alpha} = (\alpha+\beta)\left(\mathbf{I} + \frac{\alpha\beta}{2} d\mathbf{Z}\, d\mathbf{Z}\right),$$

where the first-order term of $d\mathbf{Z}$ disappears and the higher-order terms are neglected.
Since

$$\det\left(\mathbf{I} + \frac{\alpha\beta}{2} d\mathbf{Z}\, d\mathbf{Z}\right) = 1 + \frac{\alpha\beta}{2} \operatorname{tr}(d\mathbf{Z}\, d\mathbf{Z}), \tag{160}$$

by taking its logarithm, we have

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P} + d\mathbf{P} \,\|\, \mathbf{P}) = \frac{1}{2} \operatorname{tr}(d\mathbf{P}\,\mathbf{P}^{-1}\, d\mathbf{P}\,\mathbf{P}^{-1}), \tag{161}$$

for any α and β.

F. Proof of the Properties of the AB Log-Det Divergence

Next we provide a proof of the properties of the AB log-det divergence. The proof will only be omitted for those properties which can be readily verified from the definition of the divergence.

1. Nonnegativity; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \geq 0, \qquad \forall\alpha,\beta \in \mathbb{R}. \tag{162}$$

The proof of this property is presented in Appendix D.

2. Identity of indiscernibles; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = 0 \text{ if and only if } \mathbf{P} = \mathbf{Q}. \tag{163}$$

See Appendix D for its proof.

3. Continuity and smoothness of $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ as a function of $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$, including the singular cases when $\alpha = 0$ or $\beta = 0$, and when $\alpha = -\beta$ (see Figure 2).

4. The divergence can be explicitly expressed in terms of $\mathbf{\Lambda} = \mathrm{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, the diagonal matrix with the eigenvalues of $\mathbf{Q}^{-1}\mathbf{P}$; in the form

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\mathbf{\Lambda}\|\mathbf{I}). \tag{164}$$

Proof. From the definition of divergence and taking into account the eigenvalue decomposition $\mathbf{P}\mathbf{Q}^{-1} = \mathbf{V}\mathbf{\Lambda}\,\mathbf{V}^{-1}$, we can write

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) &= \frac{1}{\alpha\beta} \log\det \frac{\alpha\, \mathbf{V}\mathbf{\Lambda}^{\beta}\, \mathbf{V}^{-1} + \beta\, \mathbf{V}\mathbf{\Lambda}^{-\alpha}\, \mathbf{V}^{-1}}{\alpha+\beta} \\ &= \frac{1}{\alpha\beta} \log\left[\det\mathbf{V}\ \det\frac{\alpha\mathbf{\Lambda}^{\beta} + \beta\mathbf{\Lambda}^{-\alpha}}{\alpha+\beta}\ \det\mathbf{V}^{-1}\right] \\ &= \frac{1}{\alpha\beta} \log\det\frac{\alpha\, \mathbf{\Lambda}^{\beta} + \beta\, \mathbf{\Lambda}^{-\alpha}}{\alpha+\beta} &(165) \\ &= D_{AB}^{(\alpha,\beta)}(\mathbf{\Lambda}\|\mathbf{I}). &(166) \end{aligned}$$

□

5. Scaling invariance; given by

$$D_{AB}^{(\alpha,\beta)}(c\mathbf{P}\|c\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}), \tag{167}$$

for any $c > 0$.

6. For a given α and β and nonzero scaling factor $\omega \neq 0$, we have

$$D_{AB}^{(\omega\,\alpha,\,\omega\,\beta)}(\mathbf{P}\|\mathbf{Q}) = \frac{1}{\omega^2} D_{AB}^{(\alpha,\beta)}((\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})^{\omega}\|\mathbf{I})\ . \tag{168}$$

Proof. From the definition of divergence, we write

$$\begin{aligned} D_{AB}^{(\omega\,\alpha,\,\omega\,\beta)}(\mathbf{P}\|\mathbf{Q}) &= \frac{1}{(\omega\alpha)(\omega\beta)}\log\det\frac{\omega\alpha\;\mathbf{\Lambda}^{\omega\beta}+\omega\beta\;\mathbf{\Lambda}^{-\omega\alpha}}{(\omega\alpha+\omega\beta)} && (169)\\ &= \frac{1}{\omega^2}\frac{1}{\alpha\beta}\log\det\frac{\alpha\;(\mathbf{\Lambda}^{\omega})^{\beta}+\beta\;(\mathbf{\Lambda}^{\omega})^{-\alpha}}{(\alpha+\beta)} && (170)\\ &= \frac{1}{\omega^2}D_{AB}^{(\alpha,\beta)}((\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})^{\omega}\|\mathbf{I}) && (171)\end{aligned}$$

Hence, the additional inequality

$$D_{AB}^{(\alpha,\beta)}((\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})^{\omega}\|\mathbf{I}) \;\leq\; D_{AB}^{(\omega\,\alpha,\,\omega\,\beta)}(\mathbf{P}\|\mathbf{Q}) \qquad (172)$$

is obtained for $|\omega|\leq 1$. □

7. Dual-invariance under inversion (for $\omega=-1$); given by

$$D_{AB}^{(-\alpha,-\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\mathbf{P}^{-1}\|\mathbf{Q}^{-1})\,. \qquad (173)$$

8. Dual symmetry; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) = D_{AB}^{(\beta,\alpha)}(\mathbf{Q}\|\mathbf{P})\,. \qquad (174)$$

9. Affine invariance (invariance under congruence transformations); given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{A}\mathbf{P}\mathbf{A}^T\|\mathbf{A}\mathbf{Q}\mathbf{A}^T) \;=\; D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}), \qquad (175)$$

for any nonsingular matrix $\mathbf{A}\in\mathbb{R}^{n\times n}$.

Proof.

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\mathbf{A}\mathbf{P}\mathbf{A}^T\|\mathbf{A}\mathbf{Q}\mathbf{A}^T) &= \frac{1}{\alpha\beta}\log\det\frac{\alpha\;((\mathbf{A}\mathbf{P}\mathbf{A}^T)(\mathbf{A}\mathbf{Q}\mathbf{A}^T)^{-1})^{\beta}+\beta\;((\mathbf{A}\mathbf{P}\mathbf{A}^T)(\mathbf{A}\mathbf{Q}\mathbf{A}^T)^{-1})^{-\alpha}}{\alpha+\beta}\\ &= \frac{1}{\alpha\beta}\log\det\frac{\alpha\;(\mathbf{A}(\mathbf{P}\mathbf{Q}^{-1})\mathbf{A}^{-1})^{\beta}+\beta\;(\mathbf{A}(\mathbf{P}\mathbf{Q}^{-1})\mathbf{A}^{-1})^{-\alpha}}{\alpha+\beta}\\ &= \frac{1}{\alpha\beta}\log\left[\det(\mathbf{A}\mathbf{V})\;\det\frac{\alpha\mathbf{\Lambda}^{\beta}+\beta\mathbf{\Lambda}^{-\alpha}}{\alpha+\beta}\;\det(\mathbf{A}\mathbf{V})^{-1}\right]\\ &= \frac{1}{\alpha\beta}\log\det\frac{\alpha\;\mathbf{\Lambda}^{\beta}+\beta\;\mathbf{\Lambda}^{-\alpha}}{\alpha+\beta}\\ &= D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) && (176)\end{aligned}$$

□

10. Divergence lower-bound; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{X}^T\mathbf{P}\mathbf{X}\|\mathbf{X}^T\mathbf{Q}\mathbf{X}) \;\leq\; D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}), \qquad (177)$$

for any full-column rank matrix $\mathbf{X}\in\mathbb{R}^{n\times m}$ with $n\leq m$.

This result has been already proved for some special cases of α and β, especially these that lead to the S-divergence and the Riemannian metric [6]. Next, we present a different argument to prove it for any $\alpha,\beta\in\mathbb{R}$.

Proof. As already discussed, the divergence $D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})$ depends on the generalized eigenvalues of the matrix pencil $(\mathbf{P},\mathbf{Q})$, which have been denoted by $\lambda_i, i=1,\ldots,n$. Similarly, the presumed lower-bound $D_{AB}^{(\alpha,\beta)}(\mathbf{X}^T\mathbf{P}\mathbf{X}\|\mathbf{X}^T\mathbf{Q}\mathbf{X})$ is determined by $\mu_i, i = 1,\ldots,m$, the eigenvalues of the matrix pencil $(\mathbf{X}^T\mathbf{P}\mathbf{X},\mathbf{X}^T\mathbf{Q}\mathbf{X})$. Assuming that both sets of eigenvalues are arranged in decreasing order, the Cauchy interlacing inequalities [29] provide the following upper and lower-bounds for μ_j in terms of the eigenvalues of the first matrix pencil,

$$\lambda_j \leq \mu_j \leq \lambda_{n-m+j}. \tag{178}$$

We classify the eigenvalues μ_j on three sets S_μ^-, S_μ^0 and S_μ^+, according to the sign of (μ_j-1). By the affine invariance we can write

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\mathbf{X}^T\mathbf{P}\mathbf{X}\|\mathbf{X}^T\mathbf{Q}\mathbf{X}) &= D_{AB}^{(\alpha,\beta)}((\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1/2}\mathbf{X}^T\mathbf{P}\mathbf{X}(\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1/2}\|\mathbf{I}) && (179)\\ &= \sum_{\mu_j\in S_\mu^-} D_{AB}^{(\alpha,\beta)}(\mu_j\|1) + \sum_{\mu_j\in S_\mu^+} D_{AB}^{(\alpha,\beta)}(\mu_j\|1), && (180)\end{aligned}$$

where the eigenvalues $\mu_j \in S_\mu^0$ have been excluded since for them $D_{AB}^{(\alpha,\beta)}(\mu_j\|1)=0$.

With the help of (178), the first group of eigenvalues $\mu_j \in S_\mu^-$ (which are smaller than one) are one-to-one mapped with their lower-bounds λ_j, which we include in the set S_λ^-. Also those $\mu_j \in S_\mu^+$ (which are greater than one) are mapped with their upper-bounds λ_{n-m+j}, which we group in S_λ^+. It is shown in Appendix D that the scalar divergence $D_{AB}^{(\alpha,\alpha)}(\lambda\|1)$ is strictly monotone descending for $\lambda<1$, zero for $\lambda=1$ and strictly monotone ascending for $\lambda>1$. This allows one to upperbound (180) as follows

$$\begin{aligned} \sum_{\mu_j\in S_\mu^-} D_{AB}^{(\alpha,\beta)}(\mu_j\|1) + \sum_{\mu_j\in S_\mu^+} D_{AB}^{(\alpha,\beta)}(\mu_j\|1) &\leq \sum_{\lambda_j\in S_\lambda^-} D_{AB}^{(\alpha,\beta)}(\lambda_j\|1) + \sum_{\lambda_j\in S_\lambda^+} D_{AB}^{(\alpha,\beta)}(\lambda_j\|1) \\ &\leq \sum_{j=1}^{n} D_{AB}^{(\alpha,\beta)}(\lambda_j\|1) && (181)\\ &= D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}), && (182)\end{aligned}$$

obtaining the desired property. □

11. Scaling invariance under the Kronecker product; given by

$$D_{AB}^{(\alpha,\beta)}(\mathbf{Z}\otimes\mathbf{P}\|\mathbf{Z}\otimes\mathbf{Q}) = n\, D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q}) \,, \tag{183}$$

for any symmetric and positive definite matrix $\mathbf{Z}$ of rank n.

Proof. This property was obtained in [6] for the S-divergence and the Riemannian metric. With the help of the properties of the Kronecker product of matrices, the desired equality is obtained:

$$\begin{aligned} D_{AB}^{(\alpha,\beta)}(\mathbf{Z}\otimes\mathbf{P}\|\mathbf{Z}\otimes\mathbf{Q}) &= \frac{1}{\alpha\beta}\log\det\left[\frac{\alpha\left((\mathbf{Z}\otimes\mathbf{P})(\mathbf{Z}\otimes\mathbf{Q})^{-1}\right)^{\beta}+\beta\left((\mathbf{Z}\otimes\mathbf{Q})(\mathbf{Z}\otimes\mathbf{P})^{-1}\right)^{\alpha}}{\alpha+\beta}\right] \\ &= \frac{1}{\alpha\beta}\log\det\left[\frac{\alpha\,(\mathbf{I}\otimes\mathbf{P}\mathbf{Q}^{-1})^{\beta}+\beta\,(\mathbf{I}\otimes\mathbf{Q}\mathbf{P}^{-1})^{\alpha}}{\alpha+\beta}\right] \qquad (184) \\ &= \frac{1}{\alpha\beta}\log\det\left[\mathbf{I}\otimes\frac{\alpha\,(\mathbf{P}\mathbf{Q}^{-1})^{\beta}+\beta\,(\mathbf{Q}\mathbf{P}^{-1})^{\alpha}}{\alpha+\beta}\right] \qquad (185) \\ &= \frac{1}{\alpha\beta}\log\det\left[\frac{\alpha\,(\mathbf{P}\mathbf{Q}^{-1})^{\beta}+\beta\,(\mathbf{Q}\mathbf{P}^{-1})^{\alpha}}{\alpha+\beta}\right]^{n} \qquad (186) \\ &= n\,D_{AB}^{(\alpha,\beta)}(\mathbf{P}\|\mathbf{Q})\,. \qquad (187) \end{aligned}$$

□

12. Double Sided Orthogonal Procrustes property. Consider an orthogonal matrix $\boldsymbol{\Omega}\in\mathcal{O}(n)$ and two symmetric positive definite matrices $\mathbf{P}$ and $\mathbf{Q}$, with respective eigenvalue matrices $\boldsymbol{\Lambda}_{\mathbf{P}}$ and $\boldsymbol{\Lambda}_{\mathbf{Q}}$, which elements are sorted in descending order. The AB log-det divergence between $\boldsymbol{\Omega}^T\mathbf{P}\boldsymbol{\Omega}$ and $\mathbf{Q}$ is globally minimized when their eigenspaces are aligned, i.e.,

$$\min_{\boldsymbol{\Omega}\in\mathcal{O}(n)} D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Omega}^T\mathbf{P}\boldsymbol{\Omega}\|\mathbf{Q}) = D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Lambda}_{\mathbf{P}}\|\boldsymbol{\Lambda}_{\mathbf{Q}}). \qquad (188)$$

Proof. Let $\boldsymbol{\Lambda}$ denote the matrix of eigenvalues of $\boldsymbol{\Omega}^T\mathbf{P}\boldsymbol{\Omega}\mathbf{Q}^{-1}$ with its elements sorted in descending order. We start showing that for $\boldsymbol{\Delta}=\log\boldsymbol{\Lambda}$, the function $D_{AB}^{(\alpha,\beta)}(\exp\boldsymbol{\Delta}\|\mathbf{I})$ is convex. Its Hessian matrix is diagonal and positive definite, *i.e.*, with non-negative diagonal elements

$$\frac{\partial^2 D_{AB}^{(\alpha,\beta)}(e^{\Delta_{ii}}\|1)}{\partial\Delta_{ii}^2} > 0, \qquad (189)$$

where

$$\frac{\partial^2 D_{AB}^{(\alpha,\beta)}(e^{\Delta_{ii}}\|1)}{\partial\Delta_{ii}^2} = \begin{cases} \left(\frac{\beta}{\alpha+\beta}e^{-\frac{\alpha+\beta}{2}\Delta_{ii}}+\frac{\alpha}{\alpha+\beta}e^{\frac{\alpha+\beta}{2}\Delta_{ii}}\right)^{-2} & \text{for } \alpha,\beta,\alpha+\beta\neq 0 \\ e^{\beta\Delta_{ii}} & \text{for } \alpha=0 \\ (1+\alpha\Delta_{ii})^{-2} & \text{for } \alpha+\beta=0 \\ e^{\alpha\Delta_{ii}} & \text{for } \beta=0. \end{cases} \qquad (190)$$

Since $f(e^{\Delta_{ii}})=D_{AB}^{(\alpha,\beta)}(e^{\Delta_{ii}}\|1)$ is strictly convex and non-negative, we are in the conditions of the Corollary 6.15 in [47]. This result states that for two symmetric positive definite matrices $\mathbf{A}$ and $\mathbf{B}$, which vectors of eigenvalues are respectively denoted by $\vec{\boldsymbol{\lambda}}_{\mathbf{A}}^{\downarrow}$ (when sorted in descending order) and $\vec{\boldsymbol{\lambda}}_{\mathbf{B}}^{\uparrow}$ (when sorted in ascending order), the function $f(\vec{\boldsymbol{\lambda}}_{\mathbf{A}}^{\downarrow}\vec{\boldsymbol{\lambda}}_{\mathbf{B}}^{\uparrow})$ is submajorized by $f(\vec{\boldsymbol{\lambda}}_{\mathbf{AB}}^{\downarrow})$. By choosing $\mathbf{A}=\boldsymbol{\Omega}^T\mathbf{P}\boldsymbol{\Omega}$, $\mathbf{B}=\mathbf{Q}^{-1}$, and applying the corollary, we obtain

$$D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Lambda}_{\mathbf{P}}\|\boldsymbol{\Lambda}_{\mathbf{Q}}) = D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Lambda}_{\mathbf{P}}\boldsymbol{\Lambda}_{\mathbf{Q}}^{-1}\|\mathbf{I}) \le D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Lambda}\|\mathbf{I}) = D_{AB}^{(\alpha,\beta)}(\boldsymbol{\Omega}^T\mathbf{P}\boldsymbol{\Omega}\|\mathbf{Q}), \qquad (191)$$

where the equality is only reached when the eigendecompositions of the matrices $\mathbf{\Omega}^T\mathbf{P}\mathbf{\Omega} = \mathbf{V}\mathbf{\Lambda}_\mathbf{P}\mathbf{V}^T$ and $\mathbf{Q} = \mathbf{V}\mathbf{\Lambda}_\mathbf{Q}\mathbf{V}^T$, share the same matrix of eigenvectors $\mathbf{V}$. □

13. Triangle Inequality-Metric Distance Condition, for $\alpha = \beta \in \mathbb{R}$; given by

$$\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})} \leq \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Z})} + \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{Z}\|\mathbf{Q})}\ . \tag{192}$$

Proof. The proof of this property exploits the recent result that the square root of the S-divergence

$$d_{\text{Bh}}(\mathbf{P}\|\mathbf{Q}) = \sqrt{D_S(\mathbf{P}\|\mathbf{Q})} = 2\sqrt{\log \frac{\det \frac{1}{2}(\mathbf{P}+\mathbf{Q})}{\sqrt{\det(\mathbf{P})\det(\mathbf{Q})}}}\ . \tag{193}$$

is a metric [17]. Given three arbitrary symmetric positive definite matrices $\mathbf{P}, \mathbf{Q}, \mathbf{Z}$, with common dimensions, consider the following eigenvalue decompositions

$$\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}} = \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1^T \tag{194}$$

$$\mathbf{Q}^{-\frac{1}{2}}\mathbf{Z}\mathbf{Q}^{-\frac{1}{2}} = \mathbf{V}_2\mathbf{\Lambda}_2\mathbf{V}_2^T, \tag{195}$$

and assume that the diagonal matrices $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ have the eigenvalues sorted in a descending order.

For a given value of α in the divergence, we define $\omega = 2\alpha \neq 0$ and use properties 6 and 9 (see Equations (168) and (175)) to obtain the equivalence

$$\begin{aligned}\sqrt{D_{AB}^{(\alpha,\,\alpha)}(\mathbf{P}\|\mathbf{Q})} &= \sqrt{D_{AB}^{(\omega\,0.5,\,\omega\,0.5)}(\mathbf{P}\|\mathbf{Q})} \\ &= \sqrt{\frac{1}{\omega^2}D_{AB}^{(0.5,0.5)}((\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})^{\,\omega}\|\mathbf{I})} \\ &= \frac{1}{2|\alpha|}\sqrt{D_{AB}^{(0.5,0.5)}(\mathbf{\Lambda}_1^{2\alpha}\|\mathbf{I})} \\ &= \frac{1}{2|\alpha|}d_{\text{Bh}}(\mathbf{\Lambda}_1^{2\alpha}\|\mathbf{I})\ ,\end{aligned} \tag{196}$$

Since the S-divergence satisfies the triangle inequality for diagonal matrices [5,6,17]

$$d_{\text{Bh}}(\mathbf{\Lambda}_1^{2\alpha}\|\mathbf{I}) \leq d_{\text{Bh}}(\mathbf{\Lambda}_1^{2\alpha}\|\mathbf{\Lambda}_2^{2\alpha}) + d_{\text{Bh}}(\mathbf{\Lambda}_2^{2\alpha}\|\mathbf{I}), \tag{197}$$

from (196), this implies that

$$\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})} \leq \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{\Lambda}_1\|\mathbf{\Lambda}_2)} + \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{\Lambda}_2\|\mathbf{I})} \tag{198}$$

In similarity with the proof of the metric condition for S-divergence [6], we can use property 12 to bound above the first term in the right-hand-side of the equation by

$$\begin{aligned}\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{\Lambda}_1\|\mathbf{\Lambda}_2)} &\leq \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1^T\|\mathbf{V}_2\mathbf{\Lambda}_2\mathbf{V}_2^T)} \\ &= \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}}\|\mathbf{Q}^{-\frac{1}{2}}\mathbf{Z}\mathbf{Q}^{-\frac{1}{2}})} \\ &= \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Z})},\end{aligned} \tag{199}$$

whereas the second term satisfies

$$\begin{aligned}\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{\Lambda}_2\|\mathbf{I})} &= \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{V}_2\mathbf{\Lambda}_2\mathbf{V}_2^T\|\mathbf{I})} \\ &= \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{Q}^{-\frac{1}{2}}\mathbf{P}\mathbf{Q}^{-\frac{1}{2}}\|\mathbf{I})} \\ &= \sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{Z}\|\mathbf{Q})}. \end{aligned} \tag{200}$$

After bounding the right-hand-side of (198) with the help of (199) and (200), the divergence satisfies the desired triangle inequality (192) for $\alpha \neq 0$.

On the other hand, $\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})}$ as $\alpha \to 0$ converges to the Riemannian metric

$$\sqrt{D_{AB}^{(0,0)}(\mathbf{P}\|\mathbf{Q})} = \lim_{\alpha\to 0}\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})} \tag{201}$$

$$= \|\log(\mathbf{Q}^{-1/2}\mathbf{P}\mathbf{Q}^{-1/2})\|_F \tag{202}$$

$$= d_R(\mathbf{P}\|\mathbf{Q}) \ , \tag{203}$$

which concludes the proof of the metric condition of $\sqrt{D_{AB}^{(\alpha,\alpha)}(\mathbf{P}\|\mathbf{Q})}$ for any $\alpha \in \mathbb{R}$. □

G. Proof of Theorem 3

This theorem assumes that the range spaces of the symmetric positive semidefinite matrices $\mathbf{C}_{\boldsymbol{x}}$ and $\mathbf{C}_{\boldsymbol{y}}$ are disjoint, in the sense that they only intersect at the origin, which is the most probable situation for $n \gg r$ (where n is the size of the matrices while r is their common rank). For $\rho > 0$ the regularized versions $\tilde{\mathbf{C}}_{\boldsymbol{x}}$ and $\tilde{\mathbf{C}}_{\boldsymbol{y}}$ of these matrices are full rank.

Let $\tilde{\mathbf{\Lambda}} = \operatorname{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_n)$ denote the diagonal matrix representing the n eigenvalues of the matrix pencil $(\tilde{\mathbf{C}}_{\boldsymbol{x}}, \tilde{\mathbf{C}}_{\boldsymbol{y}})$. The AB log-det divergence between the regularized matrices is equal to the divergence between $\tilde{\mathbf{\Lambda}}$ and the identity matrix of size n, *i.e.*,

$$D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{C}}_{\boldsymbol{x}}\,\|\tilde{\mathbf{C}}_{\boldsymbol{y}}) = D_{AB}^{(\alpha,\beta)}\left(\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}}\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}}\|\,\mathbf{I}_n\right) = D_{AB}^{(\alpha,\beta)}\left(\tilde{\mathbf{\Lambda}}\|\,\mathbf{I}_n\right). \tag{204}$$

The positive eigenvalues of the matrix pencil satisfy

$$\tilde{\mathbf{\Lambda}} \equiv \operatorname{diag} Eig_+\left\{(\tilde{\mathbf{C}}_{\boldsymbol{y}})^{-\frac{1}{2}}\tilde{\mathbf{C}}_{\boldsymbol{x}}(\tilde{\mathbf{C}}_{\boldsymbol{y}})^{-\frac{1}{2}})\right\} = \operatorname{diag} Eig_+\left\{\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}\right\}, \tag{205}$$

therefore, the divergence can be directly estimated from the eigenvalues of $\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}$. In order to simplify this matrix product, we first express $\tilde{\mathbf{C}}_{\boldsymbol{x}}$ and $\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}$ in term of the auxiliary matrices

$$\mathbf{T}_{\boldsymbol{x}} = \mathbf{U}_{\boldsymbol{x}}(\mathbf{\Lambda}_{\boldsymbol{x}} - \rho\mathbf{I}_r)^{\frac{1}{2}} \quad \text{and} \quad \mathbf{T}_{\boldsymbol{y}} = \mathbf{U}_{\boldsymbol{y}}(\mathbf{\Lambda}_{\boldsymbol{y}} - \rho\mathbf{I}_r)^{\frac{1}{2}}. \tag{206}$$

In this way, they are written as a scaled version of the identity matrix plus a symmetric term:

$$\begin{aligned}\tilde{\mathbf{C}}_{\boldsymbol{x}} &= \mathbf{C}_{\boldsymbol{x}} + \rho\,\mathbf{U}_{\boldsymbol{x}}^{\perp}(\mathbf{U}_{\boldsymbol{x}}^{\perp})^T \\ &= \mathbf{U}_{\boldsymbol{x}}\mathbf{\Lambda}_{\boldsymbol{x}}\mathbf{U}_{\boldsymbol{x}}^T + \rho(\mathbf{I}_n - \mathbf{U}_{\boldsymbol{x}}\mathbf{U}_{\boldsymbol{x}}^T) \\ &= \rho\mathbf{I}_n + \mathbf{U}_{\boldsymbol{x}}(\mathbf{\Lambda}_{\boldsymbol{x}} - \rho\mathbf{I}_r)\mathbf{U}_{\boldsymbol{x}}^T \\ &= \rho\mathbf{I}_n + \mathbf{T}_{\boldsymbol{x}}\mathbf{T}_{\boldsymbol{x}}^T, \end{aligned} \tag{207}$$

and

$$\begin{aligned}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1} &= \mathbf{C}_{\boldsymbol{y}}^{+} + \rho^{-1}\mathbf{U}_{\boldsymbol{y}}^{\perp}(\mathbf{U}_{\boldsymbol{y}}^{\perp})^{T} \\ &= \mathbf{U}_{\boldsymbol{y}}\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{U}_{\boldsymbol{y}}^{T} + \rho^{-1}(\mathbf{I}_{n} - \mathbf{U}_{\boldsymbol{y}}\mathbf{U}_{\boldsymbol{y}}^{T}) \\ &= \rho^{-1}\mathbf{I}_{n} - \rho^{-1}\mathbf{U}_{\boldsymbol{y}}(\boldsymbol{\Lambda}_{\boldsymbol{y}} + \rho\mathbf{I}_{r})\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{U}_{\boldsymbol{y}}^{T} \\ &= \rho^{-1}\mathbf{I}_{n} - \rho^{-1}\mathbf{T}_{\boldsymbol{y}}\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{T}_{\boldsymbol{y}}^{T}. \end{aligned} \tag{208}$$

Next, using (207) and (208), we expand the product

$$\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1} = \mathbf{I}_{n} + \rho^{-1}\mathbf{T}_{\boldsymbol{x}}\mathbf{T}_{\boldsymbol{x}}^{T}(\mathbf{I}_{n} - \mathbf{T}_{\boldsymbol{y}}\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{T}_{\boldsymbol{y}}^{T}) + \mathbf{R} \tag{209}$$

and approximate the eigenvectors $\mathbf{U}_{\boldsymbol{y}} \to \mathbf{U}_{\boldsymbol{x}}$ of the residual matrix $\mathbf{R}$ to obtain the estimate

$$\mathbf{R} \equiv -\mathbf{U}_{\boldsymbol{y}}(\mathbf{I}_{r} + \rho\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1})\mathbf{U}_{\boldsymbol{y}}^{T} \approx -\mathbf{U}_{\boldsymbol{x}}(\mathbf{I}_{r} + \rho\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1})\mathbf{U}_{\boldsymbol{x}}^{T} \equiv \hat{\mathbf{R}}. \tag{210}$$

Hence, it is not difficult to see that the estimated residual is equal to

$$\hat{\mathbf{R}} = -\mathbf{T}_{\boldsymbol{x}}(\mathbf{I}_{r} + \rho\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1})\mathbf{T}_{\boldsymbol{x}}^{+}. \tag{211}$$

After substituting (211) in (209) and collecting common terms, we obtain the expansion

$$\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1} = \underbrace{\mathbf{I}_{n} + \mathbf{T}_{\boldsymbol{x}}\left(\rho^{-1}\mathbf{T}_{\boldsymbol{x}}^{T} - \rho^{-1}\mathbf{T}_{\boldsymbol{x}}^{T}\mathbf{T}_{\boldsymbol{y}}\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{T}_{\boldsymbol{y}}^{T} - (\mathbf{I}_{r} + \rho\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1})\mathbf{T}_{\boldsymbol{x}}^{+}\right)}_{\widehat{\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}}} + \mathrm{O}(\rho^{0}). \tag{212}$$

Let $Eig_{\lessgtr 1}\{\cdot\}$ denote the arrangement of the ordered eigenvalues of the matrix argument after excluding those that are equal to 1. For convenience, we reformulate the property proved in [30] that for any pair of matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$, the non-zero eigenvalues of $\mathbf{A}\mathbf{B}^{T}$ and of $\mathbf{B}^{T}\mathbf{A}$ are the same, into the following proposition.

Proposition 1. *For any pair of $m \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$, the eigenvalues of the matrices $\mathbf{I}_{m} + \mathbf{A}\mathbf{B}^{T}$ and $\mathbf{I}_{n} + \mathbf{B}^{T}\mathbf{A}$, which are not equal to 1, coincide.*

$$Eig_{\lessgtr 1}\left\{\mathbf{I}_{m} + \mathbf{A}\mathbf{B}^{T}\right\} = Eig_{\lessgtr 1}\left\{\mathbf{I}_{n} + \mathbf{B}^{T}\mathbf{A}\right\} \tag{213}$$

Since range spaces of $\mathbf{C}_{\boldsymbol{x}}$ and of $\mathbf{C}_{\boldsymbol{y}}$ only intersect at the origin, the approximation matrix $\widehat{\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}}$ has r dominant eigenvalues of order $\mathrm{O}(\rho^{-1})$ and $(n-r)$ remaining eigenvalues equal to 1. Using Proposition 1, these r dominant eigenvalues are given by

$$\begin{aligned} Eig_{\lessgtr 1}\left\{\widehat{\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}}\right\} &= Eig_{\lessgtr 1}\left\{\mathbf{I}_{r} + \left(\rho^{-1}\mathbf{T}_{\boldsymbol{x}}^{T} - \rho^{-1}\mathbf{T}_{\boldsymbol{x}}^{T}\mathbf{T}_{\boldsymbol{y}}\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{T}_{\boldsymbol{y}}^{T} - (\mathbf{I}_{r} + \rho\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1})\mathbf{T}_{\boldsymbol{x}}^{+}\right)\mathbf{T}_{\boldsymbol{x}}\right\} \\ &= Eig_{\lessgtr 1}\left\{\rho^{-1}\mathbf{T}_{\boldsymbol{x}}^{T}\mathbf{T}_{\boldsymbol{x}} - \rho^{-1}\mathbf{T}_{\boldsymbol{x}}^{T}\mathbf{T}_{\boldsymbol{y}}\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{T}_{\boldsymbol{y}}^{T}\mathbf{T}_{\boldsymbol{x}} - \rho\boldsymbol{\Lambda}_{\boldsymbol{y}}^{-1}\right\}. \end{aligned} \tag{214}$$

Let $\tilde{\boldsymbol{\Lambda}}_{max}$ and $\tilde{\boldsymbol{\Lambda}}_{min}$, respectively denote the diagonal submatrices of $\tilde{\boldsymbol{\Lambda}}$ with the r largest and with the r smallest eigenvalues. From the definitions in (66) and (206), one can recognize that $\mathbf{T}_{\boldsymbol{x}}^{T}\mathbf{T}_{\boldsymbol{x}} = \boldsymbol{\Lambda}_{\boldsymbol{x}} - \rho\mathbf{I}_{r}$, while $\mathbf{T}_{\boldsymbol{x}}^T\mathbf{T}_{\boldsymbol{y}} = \mathbf{W}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{xy}}\mathbf{W}_{\boldsymbol{y}}$, and substituting them in (214) we obtain the estimate of the r largest eigenvalues

$$\begin{aligned}\hat{\mathbf{\Lambda}}_{\max} &= \operatorname{diag} Eig_{\leqslant 1}\left\{\widehat{\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}}\right\} && (215)\\ &= \operatorname{diag} Eig_{\leqslant 1}\{\underbrace{\rho^{-1}\mathbf{\Lambda}_{\boldsymbol{x}} - \mathbf{I} - \rho\mathbf{\Lambda}_{\boldsymbol{y}}^{-1} - \rho^{-1}\mathbf{W}_{\boldsymbol{x}}^T\mathbf{K}_{\boldsymbol{xy}}\mathbf{W}_{\boldsymbol{y}}\mathbf{\Lambda}_{\boldsymbol{y}}^{-1}\mathbf{W}_{\boldsymbol{y}}^T\mathbf{K}_{\boldsymbol{yx}}\mathbf{W}_{\boldsymbol{x}}}_{\rho^{-1}\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}^{(\rho)}}\}. && (216)\end{aligned}$$

The relative error between these eigenvalues and the r largest eigenvalues of $\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}$ is of order $\mathrm{O}(\rho)$. This is a consequence of the fact that these eigenvalues are $\mathrm{O}(\rho^{-1})$, while the Frobenius norm of the error matrix is $\mathrm{O}(\rho^0)$. Then, the relative error between the dominant eigenvalues of the two matrices can be bounded above by

$$\left(\frac{\sum_{i=1}^r(\tilde{\lambda}_i - \hat{\lambda}_i)^2}{\sum_{i=1}^r\tilde{\lambda}_i^2}\right)^{\frac{1}{2}} \leq \frac{\|\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1} - \widehat{\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-1}}\|_F}{\left(\sum_{i=1}^r\hat{\lambda}_i^2\right)^{\frac{1}{2}} + \mathrm{O}(\rho^0)} \equiv \frac{\mathrm{O}(\rho^0)}{\mathrm{O}(\rho^{-1})} \equiv \mathrm{O}(\rho). \qquad (217)$$

On the other hand, the r smallest eigenvalues of $\tilde{\mathbf{\Lambda}}$ are the reciprocal of the r dominant eigenvalues of the inverse matrix $(\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}}\tilde{\mathbf{C}}_{\boldsymbol{x}}\tilde{\mathbf{C}}_{\boldsymbol{y}}^{-\frac{1}{2}})^{-1}$, so we can estimate them using essentially the same procedure

$$\begin{aligned}\hat{\mathbf{\Lambda}}_{\min}^{-1} &= \operatorname{diag} Eig_{\leqslant 1}\left\{\widehat{\tilde{\mathbf{C}}_{\boldsymbol{y}}\tilde{\mathbf{C}}_{\boldsymbol{x}}^{-1}}\right\} && (218)\\ &= \operatorname{diag} Eig_{\leqslant 1}\{\rho^{-1}\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}^{(\rho)}\}. && (219)\end{aligned}$$

For a sufficient small value of $\rho > 0$, the dominant contribution to the AB log-det divergence comes from the r largest and r smallest eigenvalues of the matrix pencil $(\tilde{\mathbf{C}}_{\boldsymbol{x}}, \tilde{\mathbf{C}}_{\boldsymbol{y}})$, so we obtain the desired approximation

$$\begin{aligned}D_{AB}^{(\alpha,\beta)}\left(\tilde{\mathbf{\Lambda}}\|\,\mathbf{I}_n\right) &\approx D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{\Lambda}}_{\max}\,\|\,\mathbf{I}_r) + D_{AB}^{(\alpha,\beta)}(\tilde{\mathbf{\Lambda}}_{\min}\,\|\,\mathbf{I}_r) && (220)\\ &= D_{AB}^{(\alpha,\beta)}(\rho\tilde{\mathbf{\Lambda}}_{\max}\,\|\,\rho\mathbf{I}_r) + D_{AB}^{(\beta,\alpha)}(\rho\tilde{\mathbf{\Lambda}}_{\min}^{-1}\,\|\,\rho\mathbf{I}_r) && (221)\\ &\approx D_{AB}^{(\alpha,\beta)}(\rho\hat{\mathbf{\Lambda}}_{\max}\,\|\,\rho\mathbf{I}_r) + D_{AB}^{(\beta,\alpha)}(\rho\hat{\mathbf{\Lambda}}_{\min}^{-1}\,\|\,\rho\mathbf{I}_r) && (222)\\ &= D_{AB}^{(\alpha,\beta)}(\mathbf{C}_{\boldsymbol{x}|\boldsymbol{y}}^{(\rho)}\,\|\,\rho\mathbf{I}_r) + D_{AB}^{(\beta,\alpha)}(\mathbf{C}_{\boldsymbol{y}|\boldsymbol{x}}^{(\rho)}\,\|\,\rho\mathbf{I}_r). && (223)\end{aligned}$$

Moreover, as $\rho \to 0$, the relative error of this approximation also tends to zero.

H. Gamma Divergence for Multivariate Gaussian Densities

Recall that for a given quadratic function $f(\boldsymbol{x}) = -c + \boldsymbol{b}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}$, where $\mathbf{A}$ is an SPD matrix, the integral of $\exp\{f(\boldsymbol{x})\}$ with respect to $\boldsymbol{x}$ is given by

$$\int_\Omega e^{-\frac{1}{2}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{x}-c}d\boldsymbol{x} = (2\pi)^{\frac{N}{2}}\det(\mathbf{A})^{-\frac{1}{2}}\,e^{\frac{1}{2}\boldsymbol{b}^T\mathbf{A}^{-1}\boldsymbol{b}-c}. \qquad (224)$$

This formula is obtained by evaluating the integral as follows:

$$\begin{aligned}\int_\Omega e^{-\frac{1}{2}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{x}-c}d\boldsymbol{x} &= e^{\frac{1}{2}\boldsymbol{b}^T\mathbf{A}^{-1}\boldsymbol{b}-c}\int_\Omega e^{-\frac{1}{2}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}+\boldsymbol{b}^T\boldsymbol{x}-\frac{1}{2}\boldsymbol{b}^T\mathbf{A}^{-1}\boldsymbol{b}}d\boldsymbol{x} && (225)\\ &= e^{\frac{1}{2}\boldsymbol{b}^T\mathbf{A}^{-1}\boldsymbol{b}-c}\int_\Omega e^{(\boldsymbol{x}-\mathbf{A}^{-1}\boldsymbol{b})^T\mathbf{A}(\boldsymbol{x}-\mathbf{A}^{-1}\boldsymbol{b})}d\boldsymbol{x} && (226)\\ &= e^{\frac{1}{2}\boldsymbol{b}^T\mathbf{A}^{-1}\boldsymbol{b}-c}\,(2\pi)^{\frac{N}{2}}\det(\mathbf{A})^{-\frac{1}{2}}, && (227)\end{aligned}$$

assuming that $\mathbf{A}$ is an SPD matrix, which assures the convergence of the integral and the validity of (224).

The Gamma divergence involves the a product of densities. In the multivariate Gaussian case, this simplifies as

$$\begin{aligned} p^{\alpha}(\boldsymbol{x})q^{\beta}(\boldsymbol{x}) &= (2\pi)^{-\frac{N}{2}(\alpha+\beta)} \det(\mathbf{P})^{-\frac{\alpha}{2}} \det(\mathbf{Q})^{-\frac{\beta}{2}} \times \\ & \quad \exp\left\{-\frac{\alpha}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\mathbf{P}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - \frac{\beta}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\mathbf{Q}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \qquad (228) \\ &= d\ \exp\left\{-c + \boldsymbol{b}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\mathbf{A}\boldsymbol{x}\right\}, \qquad (229) \end{aligned}$$

where

$$\begin{aligned} \mathbf{A} &= \alpha\mathbf{P}^{-1} + \beta\mathbf{Q}^{-1}, & (230) \\ \boldsymbol{b} &= \left(\boldsymbol{\mu}_1^T\alpha\mathbf{P}^{-1} + \boldsymbol{\mu}_2^T\beta\mathbf{Q}^{-1}\right)^T, & (231) \\ c &= \frac{1}{2}\boldsymbol{\mu}_1(\alpha\mathbf{P}^{-1})\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2(\beta\mathbf{Q}^{-1})\boldsymbol{\mu}_2, & (232) \\ d &= (2\pi)^{-\frac{N}{2}(\alpha+\beta)} \det(\mathbf{P})^{-\frac{\alpha}{2}} \det(\mathbf{Q})^{-\frac{\beta}{2}}. & (233) \end{aligned}$$

Integrating this product with the help of (224), we obtain

$$\begin{aligned} \int_{\Omega} p^{\alpha}(\boldsymbol{x})q^{\beta}(\boldsymbol{x})d\boldsymbol{x} &= d\,(2\pi)^{\frac{N}{2}} \det(\mathbf{A})^{-\frac{1}{2}}\, e^{\frac{1}{2}\boldsymbol{b}^T\mathbf{A}^{-1}\boldsymbol{b}-c} \qquad (234) \\ &= (2\pi)^{\frac{N}{2}(1-(\alpha+\beta))} \det(\mathbf{P})^{-\frac{\alpha}{2}} \det(\mathbf{Q})^{-\frac{\beta}{2}}\ \det(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-\frac{1}{2}} \times \\ & \quad e^{\frac{1}{2}\left(\boldsymbol{\mu}_1^T\alpha\mathbf{P}^{-1}+\boldsymbol{\mu}_2^T\beta\mathbf{Q}^{-1}\right)(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-1}\left(\boldsymbol{\mu}_1^T\alpha\mathbf{P}^{-1}+\boldsymbol{\mu}_2^T\beta\mathbf{Q}^{-1}\right)^T} \times \\ & \quad e^{-\frac{1}{2}\boldsymbol{\mu}_1(\alpha\mathbf{P}^{-1})\boldsymbol{\mu}_1-\frac{1}{2}\boldsymbol{\mu}_2(\beta\mathbf{Q}^{-1})\boldsymbol{\mu}_2}, \qquad (235) \end{aligned}$$

provided that $\alpha\mathbf{P}^{-1} + \beta\mathbf{Q}^{-1}$ is positive definite.

Rearranging the expression in terms of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ yields

$$\begin{aligned} \int_{\Omega} p^{\alpha}(\boldsymbol{x})q^{\beta}(\boldsymbol{x})d\boldsymbol{x} &= (2\pi)^{\frac{N}{2}(1-(\alpha+\beta))} \det(\mathbf{P})^{-\frac{\alpha}{2}} \det(\mathbf{Q})^{-\frac{\beta}{2}}\ \det(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-\frac{1}{2}} \times \\ & \quad e^{\frac{1}{2}\boldsymbol{\mu}_1^T\left[\alpha\mathbf{P}^{-1}(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-1}\alpha\mathbf{P}^{-1}-\alpha\mathbf{P}^{-1}\right]\boldsymbol{\mu}_1} \times \\ & \quad e^{\frac{1}{2}\boldsymbol{\mu}_2^T\left[\beta\mathbf{Q}^{-1}(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-1}\beta\mathbf{Q}^{-1}-\alpha\mathbf{Q}^{-1}\right]\boldsymbol{\mu}_2} \times \\ & \quad e^{\boldsymbol{\mu}_1^T\alpha\mathbf{P}^{-1}(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-1}\beta\mathbf{Q}^{-1}\boldsymbol{\mu}_2}. \qquad (236) \end{aligned}$$

With the help of the Woodbury matrix identity, we simplify

$$e^{\frac{1}{2}\boldsymbol{\mu}_1^T\left[\alpha\mathbf{P}^{-1}(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-1}\alpha\mathbf{P}^{-1}-\alpha\mathbf{P}^{-1}\right]\boldsymbol{\mu}_1} = e^{-\frac{1}{2}\boldsymbol{\mu}_1^T(\alpha^{-1}\mathbf{P}+\beta^{-1}\mathbf{Q})^{-1}\boldsymbol{\mu}_1}, \qquad (237)$$

$$e^{\frac{1}{2}\boldsymbol{\mu}_2^T\left[\beta\mathbf{Q}^{-1}(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-1}\beta\mathbf{Q}^{-1}-\beta\mathbf{Q}^{-1}\right]\boldsymbol{\mu}_2} = e^{-\frac{1}{2}\boldsymbol{\mu}_2^T(\alpha^{-1}\mathbf{P}+\beta^{-1}\mathbf{Q})^{-1}\boldsymbol{\mu}_2}, \qquad (238)$$

$$e^{\boldsymbol{\mu}_1^T\alpha\mathbf{P}^{-1}(\alpha\mathbf{P}^{-1}+\beta\mathbf{Q}^{-1})^{-1}\beta\mathbf{Q}^{-1}\boldsymbol{\mu}_2} = e^{\boldsymbol{\mu}_1^T(\alpha^{-1}\mathbf{P}+\beta^{-1}\mathbf{Q})^{-1}\boldsymbol{\mu}_2}, \qquad (239)$$

and hence, arriving at the desired result:

$$\begin{aligned}\int_\Omega p^\alpha(\boldsymbol{x}) q^\beta(\boldsymbol{x}) d\boldsymbol{x} &= (2\pi)^{\frac{N}{2}(1-(\alpha+\beta))} \det(\mathbf{P})^{-\frac{\alpha}{2}} \det(\mathbf{Q})^{-\frac{\beta}{2}} \, (\alpha+\beta)^{-\frac{N}{2}} \times \\ &\quad \det\left(\frac{\alpha}{\alpha+\beta}\mathbf{P}^{-1} + \frac{\beta}{\alpha+\beta}\mathbf{Q}^{-1}\right)^{-\frac{1}{2}} \times \\ &\quad e^{-\frac{\alpha\beta}{2(\alpha+\beta)}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\left(\frac{\beta}{\alpha+\beta}\mathbf{P}+\frac{\alpha}{\alpha+\beta}\mathbf{Q}\right)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).}\end{aligned} \tag{240}$$

This formula can be can easily particularized to evaluate the integrals

$$\begin{aligned}\int_\Omega p^{\alpha+\beta}(\boldsymbol{x}) d\boldsymbol{x} &= \int_\Omega p^\alpha(\boldsymbol{x}) p^\beta(\boldsymbol{x}) d\boldsymbol{x} \\ &= (2\pi)^{\frac{N}{2}(1-(\alpha+\beta))} \det(\mathbf{P})^{-\frac{\alpha}{2}} \det(\mathbf{P})^{-\frac{\beta}{2}} \, \det(\alpha\mathbf{P}^{-1} + \beta\mathbf{P}^{-1})^{-\frac{1}{2}} \times \\ &\quad e^{-\frac{\alpha\beta}{2(\alpha+\beta)}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_1)^T\left(\frac{\beta}{\alpha+\beta}\mathbf{P}+\frac{\alpha}{\alpha+\beta}\mathbf{P}\right)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_1)} \\ &= (2\pi)^{\frac{N}{2}(1-(\alpha+\beta))}(\alpha+\beta)^{-\frac{N}{2}} \det(\mathbf{P})^{\frac{1-(\alpha+\beta)}{2}}\end{aligned} \tag{241}$$

and

$$\int_\Omega q^{\alpha+\beta}(\boldsymbol{x}) d\boldsymbol{x} = (2\pi)^{\frac{N}{2}(1-(\alpha+\beta))}(\alpha+\beta)^{-\frac{N}{2}} \det(\mathbf{Q})^{\frac{1-(\alpha+\beta)}{2}}. \tag{242}$$

By substituting these integrals into the definition of the Gamma divergence and simplifying, we obtain a generalized closed form formula:

$$\begin{aligned} D_{AC}^{(\alpha,\beta)}\left(p(\boldsymbol{x}) \| q(\boldsymbol{x})\right) &= \frac{1}{\alpha\beta} \log \frac{\left(\displaystyle\int_\Omega p^{\alpha+\beta}(\boldsymbol{x}) \, d\boldsymbol{x}\right)^{\frac{\alpha}{\alpha+\beta}} \left(\displaystyle\int_\Omega q^{\alpha+\beta}(\boldsymbol{x}) \, d\boldsymbol{x}\right)^{\frac{\beta}{\alpha+\beta}}}{\displaystyle\int_\Omega p^\alpha(\boldsymbol{x}) \, q^\beta(\boldsymbol{x}) \, d\boldsymbol{x}} \\ &= \frac{1}{2\alpha\beta} \log \frac{\det\left(\dfrac{\alpha}{\alpha+\beta}\mathbf{Q} + \dfrac{\beta}{\alpha+\beta}\mathbf{P}\right)}{\det(\mathbf{Q})^{\frac{\alpha}{\alpha+\beta}} \det(\mathbf{P})^{\frac{\beta}{\alpha+\beta}}} \\ &\quad + \frac{1}{2(\alpha+\beta)}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T \left(\frac{\alpha}{\alpha+\beta}\mathbf{Q} + \frac{\beta}{\alpha+\beta}\mathbf{P}\right)^{-1} (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2), \end{aligned} \tag{243}$$

which concludes the proof of Theorem 4.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Amari, S. Information geometry of positive measures and positive-definite matrices: Decomposable dually flat structure. *Entropy* **2014**, *16*, 2131–2145.
2. Basseville, M. Divergence measures for statistical data processing—An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633.

3. Moakher, M.; Batchelor, P.G. Symmetric Positive—Definite Matrices: From Geometry to Applications and Visualization. In *Chapter 17 in the Book: Visualization and Processing of Tensor Fields*; Weickert, J., Hagen, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; pp. 285–298.
4. Amari, S. Information geometry and its applications: Convex function and dually flat manifold. In *Emerging Trends in Visual Computing*; Nielsen, F., Ed.; Springer: Berlin/Heidelberg, Germany, 2009; pp. 75–102.
5. Chebbi, Z.; Moakher, M. Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function. *Linear Algebra Appl.* **2012**, *436*, 1872–1889.
6. Sra, S. Positive definite matrices and the S-divergence. **2013**. arXiv:1110.1773.
7. Nielsen, F.; Bhatia, R. *Matrix Information Geometry*; Springer: Berlin/Heidelberg, Germany, 2013.
8. Amari, S. Alpha-divergence is unique, belonging to both f-divergence and Bregman divergence classes. *IEEE Trans. Inf. Theory* **2009**, *55*, 4925–4931.
9. Zhang, J. Divergence function, duality, and convex analysis. *Neural Comput.* **2004**, *16*, 159–195.
10. Amari, S.; Cichocki, A. Information geometry of divergence functions. *Bull. Polish Acad. Sci.* **2010**, *58*, 183–195.
11. Cichocki, A.; Amari, S. Families of Alpha- Beta- and Gamma- divergences: Flexible and robust measures of similarities. *Entropy* **2010**, *12*, 1532–1568.
12. Cichocki, A.; Cruces, S.; Amari, S. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy* **2011**, *13*, 134–170.
13. Cichocki, A.; Zdunek, R.; Phan, A.-H.; Amari, S. *Nonnegative Matrix and Tensor Factorizations*; John Wiley & Sons Ltd.: Chichester, UK, 2009.
14. Cherian, A.; Sra, S.; Banerjee, A.; Papanikolopoulos, N. Jensen-Bregman logdet divergence with application to efficient similarity search for covariance matrices. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2161–2174.
15. Cherian, A.; Sra, S. Riemannian sparse coding for positive definite matrices. In Proceedings of the Computer Vision—ECCV 2014—13th European Conference, Zurich, Switzerland, September 6–12 2014; Volume 8691, pp. 299–314.
16. Olszewski, D.; Ster, B. Asymmetric clustering using the alpha-beta divergence. *Pattern Recognit.* **2014**, *47*, 2031–2041.
17. Sra, S. A new metric on the manifold of kernel matrices with application to matrix geometric mean. In Proceedings of the 26th Annual Conference on Neural Information Processing Systems 2012, Lake Tahoe, Nevada, USA, 3–6 December 2012; pp. 144–152.
18. Nielsen, F.; Liu, M.; Vemuri, B. Jensen divergence-based means of SPD Matrices. In *Matrix Information Geometry*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 111–122.
19. Hsieh, C.; Sustik, M.A.; Dhillon, I.; Ravikumar, P.; Poldrack, R. BIG & QUIC: Sparse inverse covariance estimation for a million variables. In Proceedings of the 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, Nevada, USA, 5–8 December 2013; pp. 3165–3173.
20. Nielsen, F.; Nock, R. A closed-form expression for the Sharma-Mittal entropy of exponential families. *CoRR* **2011**, arXiv:1112.4221v1 [cs.IT]. Available online: http://arxiv.org/abs/1112.4221 (accessed on 4 May 2015).

21. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *Multivar. Anal.* **2008**, *99*, 2053–2081.
22. Kulis, B.; Sustik, M.; Dhillon, I. Learning low-rank kernel matrices. In Proceedings of the Twenty-third International Conference on Machine Learning (ICML06), Pittsburgh, PA, USA, 25–29 July 2006; pp. 505–512.
23. Cherian, A.; Sra, S.; Banerjee, A.; Papanikolopoulos, N. Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence. In Proceedings of the IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, 6–13 November 2011; pp. 2399–2406.
24. Österreicher, F. Csiszár's f-divergences-basic properties. *RGMIA Res. Rep. Collect.* **2002**. Available online: http://rgmia.vu.edu.au/monographs/csiszar.htm (accessed on 6 May 2015).
25. Cichocki, A.; Zdunek, R.; Amari, S. Csiszár's divergences for nonnegative matrix factorization: Family of new algorithms. In *Independent Component Analysis and Blind Signal Separation*, Proceedings of 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2006), Charleston, SC, USA, 5–8 March 2006; Lecture Notes in Computer Sciences, Volume 3889; pp. 32–39.
26. Reeb, D.; Kastoryano, M.J.; Wolf, M.M. Hilbert's projective metric in quantum information theory. *J. Math. Phys.* **2011**, *52*, 082201.
27. Kim, S.; Kim, S.; Lee, H. Factorizations of invertible density matrices. *Linear Algebra Appl.* **2014**, *463*, 190–204.
28. Bhatia, R. *Positive Definite Matrices*; Princeton University Press: Princeton, NJ, USA, 2009.
29. Li, R.-C. *Rayleigh Quotient Based Optimization Methods For Eigenvalue Problems*; Summary of Lectures Delivered at Gene Golub SIAM Summer School 2013; Fudan University: Shanghai, China, 2013.
30. De Moor, B.L.R. *On the Structure and Geometry of the Product Singular Value Decomposition*; Numerical Analysis Project NA-89-06; Stanford University: Stanford, CA, USA, 1989; pp. 1–52.
31. Golub, G.H.; van Loan, C.F. *Matrix Computations*, 3rd ed.; Johns Hopkins University Press: Baltimore, MD, USA, 1996; pp. 555–571.
32. Zhou, S.K.; Chellappa, R. From Sample Similarity to Ensemble Similarity: Probabilistic Distance Measures in Reproducing Kernel Hilbert Space. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 917–929.
33. Harandi, M.; Salzmann, M.; Porikli, F. Bregman Divergences for Infinite Dimensional Covariance Matrices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014 ; pp. 1003–1010.
34. Minh, H.Q.; Biagio, M.S.; Murino, V. Log-Hilbert-Schmidt metric between positive definite operators on Hilbert spaces. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 388–396.
35. Josse, J.; Sardy, S. Adaptive Shrinkage of singular values. **2013**, arXiv:1310.6602.
36. Donoho, D.L.; Gavish, M.; Johnstone, I.M. Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model. **2013**, arXiv:1311.0851.
37. Gavish, M.; Donoho, D. Optimal shrinkage of singular values. **2014**, arXiv:1405.7511.

38. Davis, J.; Dhillon, I. Differential entropic clustering of multivariate gaussians. In Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2006; pp. 337–344.
39. Abou-Moustafa, K.; Ferrie, F. Modified divergences for Gaussian densities. In Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition, Hiroshima, Japan, 7–9 November 2012; pp. 426–436.
40. Burbea, J.; Rao, C. Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. *J. Multi. Anal.* **1982**, *12*, 575–596.
41. Hosseini, R.; Sra, S.; Theis, L.; Bethge, M. Statistical inference with the Elliptical Gamma Distribution. **2014**, arXiv:1410.4812
42. Manceur, A.; Dutilleul, P. Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **2013**, *239*, 37–49.
43. Akdemir, D.; Gupta, A. Array variate random variables with multiway Kronecker delta covariance matrix structure. *J. Algebr. Stat.* **2011**, *2*, 98–112.
44. PHoff, P.D. Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **2011**, *6*, 179–196.
45. Gerard, D.; Hoff, P. Equivariant minimax dominators of the MLE in the array normal model. **2014**, arXiv:1408.0424
46. Ohlson, M.; Ahmad, M.; von Rosen, D. The Multilinear Normal Distribution: Introduction and Some Basic Properties. *J. Multivar. Anal.* **2013**, *113*, 37–47.
47. Ando, T. Majorization, doubly stochastic matrices, and comparison of eigenvalues. *Linear Algebra Appl.* **1989**, *118*, 163–248.