HAL
open science

Diagonality measures of Hermitian positive-definite matrices with application to the approximate joint diagonalization problem

Khaled Alyani, Marco Congedo, Maher Moakher

To cite this version:

Khaled Alyani, Marco Congedo, Maher Moakher. Diagonality measures of Hermitian positive-definite matrices with application to the approximate joint diagonalization problem. Linear Algebra and its Applications, 2017, 528, pp.290 - 320. 10.1016/j.laa.2016.08.031 . hal-01578462

HAL Id: hal-01578462

https://hal.science/hal-01578462v1

Submitted on 29 Aug 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1608.06613v2 [cs.IT] 4 Sep 2016

Diagonality Measures of Hermitian Positive-Definite Matrices with Application to the Approximate Joint Diagonalization Problem

Khaled Alyani[a], Marco Congedo[b], Maher Moakher[a,*]

[a] *LAMSIN-ENIT, University of Tunis El Manar, B.P. 37, 1002 Tunis-Belvédère, Tunisia*
[b] *GIPSA-lab, CNRS and Grenoble University, Grenoble, France*

Abstract

In this paper, we introduce properly-invariant diagonality measures of Hermitian positive-definite matrices. These diagonality measures are defined as distances or divergences between a given positive-definite matrix and its diagonal part. We then give closed-form expressions of these diagonality measures and discuss their invariance properties. The diagonality measure based on the log-determinant α-divergence is general enough as it includes a diagonality criterion used by the signal processing community as a special case. These diagonality measures are then used to formulate minimization problems for finding the approximate joint diagonalizer of a given set of Hermitian positive-definite matrices. Numerical computations based on a modified Newton method are presented and commented.

Keywords: Diagonality measure, positive-definite matrices, Riemannian distance, Bhattacharyya distance, Kullback-Leibler divergence, log-det α-divergence, approximate joint diagonalization

2010 MSC: 15B48, 94A17

1. Introduction

In the last three decades, the problem of approximate joint diagonalization (AJD) of a collection of Hermitian positive-definite matrices has attracted an increasing attention from researchers within the statistics, signal processing and applied mathematics communities. In statistics, its is used to solve the common

*Corresponding author
Email addresses: khaled.alyani@yahoo.de (Khaled Alyani), marco.congedo@gipsa-lab.grenoble-inp.fr (Marco Congedo), maher.moakher@gmail.com (Maher Moakher)
URL: http://sites.google.com/site/moakhersite/ (Maher Moakher)

principal component (CPC) problem [1, 2]. In signal processing it is a fundamental tool for blind source separation (BSS) [3] and blind beamforming [4, 5]. Applied mathematicians and numerical analysts are concerned with efficient numerical algorithms for solving this problem [6, 7, 8, 9, 10, 11, 12, 13].

Given a set of $n \times n$ Hermitian positive-definite matrices $\boldsymbol{M}_k$, $k = 1, \ldots, K$ that are supposed to be diagonalized under congruence by a common invertible matrix $\boldsymbol{C}$, i.e., such that

$$\boldsymbol{D}_k = \boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H, \quad k = 1, \ldots, K,$$

where $\boldsymbol{D}_k$ are diagonal matrices, or equivalently, such that $\boldsymbol{C}^{-1}\boldsymbol{D}_k\boldsymbol{C}^{-H} = \boldsymbol{M}_k$, $k = 1, \ldots, K$, the problem of exact joint diagonalizer is to find a common diagonalizer for the given $\boldsymbol{M}_k$, $k = 1, \ldots, K$. We note that the joint diagonalizer can only be determined up to a row-wise permutation and scaling. In fact, if $\boldsymbol{C}$ is a common diagonalizer of $\boldsymbol{M}_k$, $k = 1, \ldots, K$ then so is $\boldsymbol{PDC}$ for any invertible diagonal matrix $\boldsymbol{D}$ and any permutation matrix $\boldsymbol{P}$.

In practice, due to the presence of noise and measurement errors, the given (observed) matrices $\boldsymbol{M}_k$, $k = 1, \ldots, K$ can not be simultaneously diagonalized for $K > 2$. The problem of approximate joint diagonalizer for the given $\boldsymbol{M}_k$, $k = 1, \ldots, K$ is then to find an invertible matrix $\boldsymbol{C}$ such that $\boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H$, $k = 1, \ldots, K$ are as close to diagonal matrices as possible [7, 8, 6, 9, 10, 11, 12, 13]. Hence, one has to define a measure of diagonality and then to devise algorithms that minimize a cost function which is the sum of the diagonality measures of all $\boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H$.

Since the joint diagonalizer is not unique, but arbitrary with respect to permutations and positive diagonal scalings, it is desirable that the cost function used for computing the approximate joint diagonalizer be invariant with respect to permutations and positive diagonal scalings. The overwhelming majority of the proposed approximate joint diagonalization algorithms use as cost function the Frobenius off squared norm, which is not invariant by diagonal scaling. Furthermore, this criterion is not specific to positive Hermitian matrices. The reason why this cost function has been heavily employed is because it allows simple optimization schemes. On the other hand, Pham [7] used a cost function which is invariant with respect to diagonal scalings and permutations and is specifically conceived for Hermitian positive-definite matrices. He used a diagonality measure derived from the Kullback-Leibler divergence to minimize an unconstrained optimization problem, in contrast to most joint diagonalization algorithms, which minimize a constrained optimization problem. We will propose a new cost function, based on the log-det α-divergence, that measures the degree of joint diagonalization of a set of Hermitian positive-definite matrices. This cost function includes some of the cost functions used in the literature as special cases. To the best of our knowledge, this paper is the first attempt to tackle the approximate joint diagonalizer problem from the Riemannian and information geometries point of view. We then describe convergent numerical algorithms for minimizing this cost function and hence finding a single matrix $\boldsymbol{C}$ that approximately joint diagonalizes the given set of matrices.

The remainder of this paper is organized as follows. In Section 2, we set notations and gather necessary background materials from matrix analysis, differential geometry, optimization on manifolds and information theory that will be used in the sequel. In Section 3, and as a first step, we start by introducing properly-invariant diagonality measures as distances or divergences between a given positive-definite matrix and its diagonal part. Properly invariant means that all desirable invariance properties (which will be made precise) that one would expect of such measures are satisfied. For generality, we have chosen to work with Hermitian positive-definite matrices. The results in this paper remain valid for symmetric positive-definite matrices, all what is needed is to consider that all matrices are real. We then give closed-form expressions of these diagonality measures and discuss their properties. Then, in Section 4, we define the "true" diagonality measures as the distances (or divergences) between a given positive-definite matrix and its closest, with respect to the distance (divergence), diagonal matrix. Unfortunately, not all of them can be given in closed-form. For those which can not be given explicitly, we give the nonlinear matrix equation that characterize them. In Section 5 application to the problem of approximate joint diagonalization of a finite collection of Hermitian positive-definite matrices is discussed and some computations of the approximate joint diagonalizer based on these diagonality measures are presented.

2. Notations and background materials

Let $\mathcal{M}_n(\mathbb{C})$ be the set of $n \times n$ complex matrices and $\mathrm{GL}_n(\mathbb{C})$ be its subset containing only nonsingular matrices. $\mathrm{GL}_n(\mathbb{C})$ is a Lie group, i.e., a group which is also a differentiable manifold and for which the operations of group multiplication and inverse are smooth. The tangent space at the identity is called the corresponding Lie algebra and denoted by $\mathfrak{gl}_n(\mathbb{C})$. It is the space of all linear transformations in $\mathbb{C}^n$, i.e., $\mathcal{M}_n(\mathbb{C})$. In $\mathcal{M}_n(\mathbb{C})$ we shall use the Euclidean inner product, known as the Frobenius inner product, defined by

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_F := \mathrm{tr}(\boldsymbol{B}^H \boldsymbol{A}),$$

where $\mathrm{tr}(\cdot)$ stands for the trace and the superscript ^{H} denotes the conjugate transpose. The associated norm $\|\boldsymbol{A}\|_F = \langle \boldsymbol{A}, \boldsymbol{A} \rangle_F^{1/2}$ is used to define the Euclidean distance on $\mathcal{M}_n(\mathbb{C})$,

$$d_F(\boldsymbol{A}, \boldsymbol{B}) := \|\boldsymbol{A} - \boldsymbol{B}\|_F.$$

2.1. The Diag *operator*

We will denote by Diag the linear operator, from $\mathcal{M}_n(\mathbb{C})$ into itself, which assigns to each matrix $\boldsymbol{A}$ the diagonal matrix $\mathrm{Diag}(\boldsymbol{A})$ whose diagonal entries are a_{kk}, the diagonal entries of $\boldsymbol{A}$. We note that $\mathrm{Diag}(\boldsymbol{A})$ can be expressed as $\mathrm{Diag}(\boldsymbol{A}) = \boldsymbol{A} \circ \boldsymbol{I} = \boldsymbol{I} \circ \boldsymbol{A}$, where $\circ$ represents the Hadamard product (entry-wise product). It is also known as the Schur product. Then, if we represent linear

maps from $\mathcal{M}_n(\mathbb{C})$ into itself by its $n^2 \times n^2$ matrix in the Canonical basis, it is easy to see that the differential of $\mathrm{Diag}(\boldsymbol{A})$ with respect to $\boldsymbol{A}$ is given by

$$D\,\mathrm{Diag}(\boldsymbol{A}) = \mathrm{diag}(\mathrm{vec}(\boldsymbol{I})),$$

where $\mathrm{vec}(\cdot)$ stands for the operator that transforms an $n \times n$ complex-valued matrix into a vector in $\mathbb{C}^{n^2}$ by stacking the columns of the matrix one underneath the other, and $\mathrm{diag}(\cdot)$ is the operator that assigns to each vector in $\mathbb{C}^m$ an $m \times m$ diagonal matrix with the elements of the vector on its main diagonal.

Let us define the diagonal unitary matrix $\boldsymbol{U}_\theta = \mathrm{diag}(e^{i\theta}, e^{2i\theta}, \ldots, e^{ni\theta})$, where $\theta \in [0, 2\pi)$. Then, it is interesting to observe that $\mathrm{Diag}(\boldsymbol{A})$ can be represented as the continuous average of rotated versions of $\boldsymbol{A}$ [14, 15]

$$\mathrm{Diag}(\boldsymbol{A}) = \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{U}_\theta \boldsymbol{A} \boldsymbol{U}_\theta^H \; d\theta,$$

or, as the discrete average of rotated versions of $\boldsymbol{A}$

$$\mathrm{Diag}(\boldsymbol{A}) = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{U}_{\theta_k} \boldsymbol{A} \boldsymbol{U}_{\theta_k}^H, \quad \text{where } \theta_k = \frac{2(k-1)\pi}{n}.$$

So that the differential of $\mathrm{Diag}(\boldsymbol{A})$ with respect to $\boldsymbol{A}$ can also be written as

$$D\,\mathrm{Diag}(\boldsymbol{A}) = \frac{1}{2\pi}\int_0^{2\pi} \boldsymbol{U}_\theta \otimes \boldsymbol{U}_\theta \; d\theta = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{U}_{\theta_k} \otimes \boldsymbol{U}_{\theta_k},$$

where $\otimes$ stands for the Kronecker product.

2.2. Exponential and logarithms

The exponential of a matrix in $\mathfrak{gl}_n(\mathbb{C})$ is given, as usual, by the convergent series

$$\exp \boldsymbol{A} = \sum_{k=0}^{\infty} \frac{1}{k!} \boldsymbol{A}^k.$$

We remark that the product of the exponentials of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is equal to $\exp(\boldsymbol{A} + \boldsymbol{B})$ only when $\boldsymbol{A}$ and $\boldsymbol{B}$ commute.

Logarithms of $\boldsymbol{A}$ in $\mathcal{M}_n(\mathbb{C})$ are solutions of the matrix equation $\exp \boldsymbol{X} = \boldsymbol{A}$. When $\boldsymbol{A}$ does not have eigenvalues on the closed negative real line, there exists a unique logarithm, called the principal logarithm and denoted by $\log \boldsymbol{A}$, whose spectrum lies in the infinite strip $\{z \in \mathbb{C} \,:\, -\pi < \mathrm{Im}(z) < \pi\}$, of the complex plane. Furthermore, if for a given matrix norm $\|\cdot\|$ we have $\|\boldsymbol{A}\| < 1$, then the series $\sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \boldsymbol{A}^k$ converges to $\log(\boldsymbol{I} + \boldsymbol{A})$, where $\boldsymbol{I}$ denotes the identity transformation in $\mathbb{C}^n$. Therefore, one can write

$$\log(\boldsymbol{I} + \boldsymbol{A}) = \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k} \boldsymbol{A}^k.$$

We note that, in general, $\log(\boldsymbol{AB}) \neq \log \boldsymbol{A} + \log \boldsymbol{B}$. Finally, we recall the important fact:

$$f(\boldsymbol{C}^{-1}\boldsymbol{BC}) = \boldsymbol{C}^{-1} f(\boldsymbol{B})\boldsymbol{C},$$

for any invertible matrix $\boldsymbol{C}$ and any analytic function f, such as exp and log.

2.3. Geodesic convexity, gradient and Hessian

For a Riemannian manifold $\mathscr{M}$ we denote by $T_x\mathscr{M}$ the tangent space at $x \in \mathscr{M}$ and by $\exp_x : T_x\mathscr{M} \to \mathscr{M}$ the exponential mapping at x, i.e. $\exp_x(v) := \gamma(1)$ where γ is the unique geodesic with $\gamma(0) = x$ and $\gamma'(0) = v$.

A subset $\mathscr{A}$ of a Riemannian manifold $\mathscr{M}$ is said to be convex if the shortest geodesic curve between any two points x and y in $\mathscr{A}$ is unique in $\mathscr{M}$ and lies entirely in $\mathscr{A}$. A real-valued function defined on a convex subset $\mathscr{A}$ of $\mathscr{M}$ is said to be convex if its restriction to any geodesic path is convex, i.e., if $t \mapsto \hat{f}(t) := f(\exp_x(tu))$ is convex over its domain for all $x \in \mathscr{M}$ and $u \in T_x(\mathscr{M})$.

Let f be a real-valued function defined on a Riemannian manifold $\mathscr{M}$. If f is differentiable then the gradient ∇f is the unique tangent vector at x such that for any direction u in the tangent space to $\mathscr{M}$ at x, we have

$$\langle u, \nabla f\rangle = \left.\frac{d}{dt}\right|_{t=0} f(\exp_x(tu)),$$

where $\langle \cdot, \cdot \rangle$ denotes the Riemannian inner product on the tangent space. If f is twice differentiable, then the Hessian $\nabla^2 f$ of f is given by

$$\langle u, \nabla^2 f v\rangle = \left.\frac{\partial^2}{\partial t \partial s}\right|_{t=0,s=0} f(\exp_x(tu + sv)).$$

The second-order Taylor expansion of the function f at $x \in \mathscr{M}$ is therefore given by

$$f(\exp_x(v)) = f(x) + \langle \nabla f, v\rangle + \tfrac{1}{2}\langle v, \nabla^2 f v\rangle + O(|v|^3).$$

2.4. The cone of Hermitian positive-definite matrices

We denote by

$$\mathscr{H}_n := \left\{ \boldsymbol{A} \in \mathcal{M}_n(\mathbb{C}),\ \boldsymbol{A}^H = \boldsymbol{A} \right\},$$

the n^2-dimensional vector space of all $n \times n$ Hermitian matrices and denote by

$$\mathscr{H}_n^{++} := \left\{ \boldsymbol{A} \in \mathscr{H}_n,\ \boldsymbol{A} > 0 \right\},$$

the set of all $n \times n$ Hermitian positive-definite matrices. Here $\boldsymbol{A} > 0$ means that the quadratic form $\boldsymbol{x}^H \boldsymbol{A}\boldsymbol{x}$ is positive for all $\boldsymbol{x} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$. It is well known that $\mathscr{H}_n^{++}$ is an open convex cone; i.e., if $\boldsymbol{P}$ and $\boldsymbol{Q}$ are in $\mathscr{H}_n^{++}$, so is $\boldsymbol{P} + t\boldsymbol{Q}$ for any $t > 0$. Its closure is the set $\mathscr{H}_n^{+}$ of all $n \times n$ Hermitian positive semi-definite matrices and its boundary $\partial\mathscr{H}_n^{++}$ is the set of all $n \times n$ Hermitian positive semi-definite singular matrices.

We recall that the exponential map from $\mathscr{H}_n$ to $\mathscr{H}_n^{++}$ is one-to-one and onto. In other words, the exponential of any Hermitian matrix is a Hermitian

positive-definite matrix, and the inverse of the exponential (i.e., the principal logarithm) of any Hermitian positive-definite matrix is a Hermitian matrix.

As $\mathscr{H}_n^{++}$ is an open subset of $\mathscr{H}_n$, for each $\boldsymbol{P} \in \mathscr{H}_n^{++}$ we identify the set $T_{\boldsymbol{P}}$ of tangent vectors to $\mathscr{H}_n^{++}$ at $\boldsymbol{P}$ with $\mathscr{H}_n$. On the tangent space at $\boldsymbol{P}$ we define the positive-definite inner product and corresponding norm [16, 17]

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\boldsymbol{P}} = \operatorname{tr}(\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{B}), \qquad \|\boldsymbol{A}\|_{\boldsymbol{P}} = \langle \boldsymbol{A}, \boldsymbol{A} \rangle_{\boldsymbol{P}}^{1/2}, \tag{1}$$

that smoothly depend on the base point $\boldsymbol{P}$. The positive definiteness is a consequence of the positive definiteness of the Frobenius inner product for

$$\langle \boldsymbol{A}, \boldsymbol{A} \rangle_{\boldsymbol{P}} = \operatorname{tr}(\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}) = \|\boldsymbol{P}^{-1/2}\boldsymbol{A}\boldsymbol{P}^{-1/2}\|_F^2.$$

Let $[a, b]$ be a closed interval in $\mathbb{R}$, and let $\boldsymbol{\Gamma} : [a, b] \to \mathscr{H}_n^{++}$ be a sufficiently smooth curve in $\mathscr{H}_n^{++}$. We define the length of $\boldsymbol{\Gamma}$ by

$$\mathcal{L}(\boldsymbol{\Gamma}) := \int_a^b \|\dot{\boldsymbol{\Gamma}}(t)\|_{\boldsymbol{\Gamma}(\boldsymbol{t})} dt.$$

We note that the length $\mathcal{L}(\boldsymbol{\Gamma})$ is invariant under congruent transformations, i.e., $\boldsymbol{\Gamma} \mapsto \boldsymbol{C}\boldsymbol{\Gamma}\boldsymbol{C}^H$, where $\boldsymbol{C}$ is any fixed element of $\mathrm{GL}_n(\mathbb{C})$. As $\frac{d\boldsymbol{\Gamma}^{-1}}{dt} = -\boldsymbol{\Gamma}^{-1}\frac{d\boldsymbol{\Gamma}}{dt}\boldsymbol{\Gamma}^{-1}$, one can readily see that this length is also invariant under inversion.

The distance between two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ in $\mathscr{H}_n^{++}$, considered as a differentiable manifold, is the infimum of lengths of curves connecting them:

$$d_{\mathscr{H}_n^{++}}(\boldsymbol{A}, \boldsymbol{B}) := \inf \left\{ \mathcal{L}(\boldsymbol{\Gamma}) | \boldsymbol{\Gamma} : [a, b] \to \mathscr{H}_n^{++} \text{ with } \boldsymbol{\Gamma}(a) = \boldsymbol{A}, \boldsymbol{\Gamma}(b) = \boldsymbol{B} \right\}.$$

This metric makes $\mathscr{H}_n^{++}$ a Riemannian manifold of dimension n^2.

The geodesic emanating from $\boldsymbol{I}$ in the direction of $\boldsymbol{S}$, a Hermitian matrix in the tangent space, is given explicitly by $\exp(t\boldsymbol{S})$. Using invariance under congruent transformations, the geodesic $\boldsymbol{P}(t)$ such that $\boldsymbol{P}(0) = \boldsymbol{P}$ and $\dot{\boldsymbol{P}}(0) = \boldsymbol{S}$ is therefore given by

$$\boldsymbol{P}(t) = \boldsymbol{P}^{1/2} \exp(t\boldsymbol{P}^{-1/2}\boldsymbol{S}\boldsymbol{P}^{-1/2})\boldsymbol{P}^{1/2}.$$

It follows that the Riemannian distance between $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ in $\mathscr{H}_n^{++}$ is

$$d_{\mathscr{H}_n^{++}}(\boldsymbol{P}_1, \boldsymbol{P}_2) = \|\log(\boldsymbol{P}_1^{-1/2}\boldsymbol{P}_2\boldsymbol{P}_1^{-1/2})\|_F = \left[\sum_{i=1}^n \log^2 \lambda_i\right]^{1/2}, \tag{2}$$

where λ_i, $i = 1, \ldots, n$, are the (positive) eigenvalues of $\boldsymbol{P}_1^{-1}\boldsymbol{P}_2$. Even though in general $\boldsymbol{P}_1^{-1}\boldsymbol{P}_2$ is not Hermitian, its eigenvalues are real and positive. This can be seen by noting that $\boldsymbol{P}_1^{-1}\boldsymbol{P}_2$ is similar to the Hermitian positive-definite matrix $\boldsymbol{P}_1^{-1/2}\boldsymbol{P}_2\boldsymbol{P}_1^{-1/2}$. It is important to note here that the real-valued function defined on $\mathscr{H}_n^{++}$ by $\boldsymbol{P} \mapsto d_{\mathscr{H}_n^{++}}(\boldsymbol{P}, \boldsymbol{S})$, where $\boldsymbol{S} \in \mathscr{H}_n^{++}$ is fixed, is (geodesically) convex. We note in passing that $\mathscr{H}_n^{++}$ is a homogeneous space of the Lie

group $\mathrm{GL}_n(\mathbb{C})$ (by identifying $\mathscr{H}_n^{++}$ with the quotient $\mathrm{GL}_n(\mathbb{C})/U_n(\mathbb{C})$). It is also a symmetric space of noncompact type [18].

It is important here to note that the subset $\mathscr{D}_n^{++}$ of diagonal positive-definite matrices is a totally geodesic submanifold of $\mathscr{H}_n^{++}$, which is geodesically convex and closed with respect to the Riemannian metric defined by the inner product $(1)_1$.

2.5. Divergence functions

We give a formal definition of a divergence function on a Riemannian manifold.

Definition 1. *A divergence function over a Riemannian manifold $\mathscr{M}$ is a real-valued function J on the Cartesian product manifold $\mathscr{M} \times \mathscr{M}$ which satisfies the following conditions [19]:*

i) Positive definiteness: $J(x, y) \geq 0$ for all x and y in $\mathscr{M}$ with equality if and only if $x = y$.

ii) Differentiability: $J(x, y)$ is twice differentiable with respect to x and its Hessian with respect to x, when evaluated at $y = x$, is positive definite.

We note that the divergence is almost a distance function except that it needs not to be symmetric with respect to its arguments nor to satisfy the triangle inequality. For instance, the square of a distance function is a (symmetric) divergence function. In some respects, a divergence function is a generalization of squared distances and has the physical dimension (i.e., unit) of squared distance.

Let $f : \Omega \to \mathbb{R}$ be a twice differentiable and strictly convex function defined on a closed convex set Ω of a vector space V. We give below two examples of systematic ways of constructing divergence functions from f.

1. The function defined by

$$B_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle, \tag{3}$$

is a divergence function on Ω called the Bregman divergence [20]. Here $\langle \cdot, \cdot \rangle$ denotes the inner product in V.

2. For a real parameter α such that $|\alpha| < 1$, the functions

$$D_f^\alpha(x, y) = \tfrac{4}{1-\alpha^2} \left[\tfrac{1-\alpha}{2} f(x) + \tfrac{1+\alpha}{2} f(y) - f(\tfrac{1-\alpha}{2} x + \tfrac{1+\alpha}{2} y) \right], \tag{4}$$

defines a one-parameter family of divergence functions called α-divergence functions [21]. Furthermore, by taking limits in (4) we have

$$D_f^1(x, y) := \lim_{\alpha \to 1^-} D_f^\alpha(x, y) = B_f(x, y),$$
$$D_f^{-1}(x, y) := \lim_{\alpha \to -1^+} D_f^\alpha(x, y) = B_f(y, x).$$

The positive definiteness of the functions (3) and (4) is guaranteed by the strict convexity of f. Furthermore, we note that the Hessian with respect to x when evaluated at $y = x$ for both functions (3) and (4) is equal to the Hessian of f at x which is positive definite by the strict convexity hypothesis on f.

Now, if we consider the logarithmic-barrier function $f(\boldsymbol{X}) = -\log\det\boldsymbol{X}$, which is a strictly convex function defined on the convex subset $\mathscr{H}_n^{++}$ of the vector space $\mathscr{H}_n$, we obtain for $-1 \leq \alpha \leq 1$, the one-parameter family of divergence functions on the space of Hermitian positive-definite matrices called log-det α-divergence functions [22]

$$\begin{cases} D_{LD}^{\alpha}(\boldsymbol{A},\boldsymbol{B}) = \frac{4}{1-\alpha^2}\log\frac{\det\left(\frac{1-\alpha}{2}\boldsymbol{A}+\frac{1+\alpha}{2}\boldsymbol{B}\right)}{\left(\det\boldsymbol{A}\right)^{\frac{1-\alpha}{2}}\left(\det\boldsymbol{B}\right)^{\frac{1+\alpha}{2}}}, \quad -1<\alpha<1, \\ D_{LD}^{-1}(\boldsymbol{A},\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{B}-\boldsymbol{I}) - \log\det(\boldsymbol{A}^{-1}\boldsymbol{B}), \\ D_{LD}^{1}(\boldsymbol{A},\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}^{-1}\boldsymbol{A}-\boldsymbol{I}) - \log\det(\boldsymbol{B}^{-1}\boldsymbol{A}). \end{cases} \tag{5}$$

2.6. Projection onto a convex set

Let $\mathscr{S}$ be a nonempty closed convex subset of a Riemannian manifold $\mathscr{M}$. If $d(\cdot,\cdot)$ is a distance function on $\mathscr{M}$, then for any $x \in \mathscr{M}$ there exists a unique element $x_d \in \mathscr{S}$ such that $d(x, x_d) \leq d(x, y)$ for all $y \in \mathscr{S}$. The element x_d is called the metric projection of x onto $\mathscr{S}$.

Similarly, if $D(\cdot,\cdot)$ is a divergence function on $\mathscr{M}$, then for any $x \in \mathscr{M}$ there exists a unique element $x_D \in \mathscr{S}$ such that $D(x, x_D) \leq D(x, y)$ for all $y \in \mathscr{S}$ [23]. The element x_D is called the divergence projection of x onto $\mathscr{S}$.

3. Properly invariant diagonality measures

We define a measure of diagonality of a matrix as a non-negative quantity that vanishes only for diagonal matrices. One way to construct such measures is to use a distance (or a divergence) function and define the measure of diagonality as the squared distance (or the divergence) between the given matrix and its diagonal part. The minimal set of invariance properties that we would expect from a diagonality measure $\mathscr{D}(\cdot)$ are the following:

1. Invariance under congruence transformations by permutation matrices:

$$\mathscr{D}(\boldsymbol{PAP}) = \mathscr{D}(\boldsymbol{A}), \text{ for all permutation matrices } \boldsymbol{P}.$$

2. Invariance under congruence transformations by invertible diagonal matrices:

$$\mathscr{D}(\boldsymbol{DAD}) = \mathscr{D}(\boldsymbol{A}), \text{ for all invertible diagonal matrices } \boldsymbol{D}.$$

Note that Property 2 implies invariance under positive scaling, i.e., $\mathscr{D}(\mu\boldsymbol{A}) = \mathscr{D}(\boldsymbol{A})$, $\forall\mu > 0$.

The simplest diagonality measure of a given matrix $\boldsymbol{A}$ could be obtained by half the squared norm of its off-diagonal part $\boldsymbol{A} - \operatorname{Diag} \boldsymbol{A}$. For example, if we consider the Euclidean distance induced from the matrix Frobenius inner product, then the Frobenius measure of diagonality of a Hermitian positive-definite matrix $\boldsymbol{A}$ is defined as half the square of the Frobenius distance between $\boldsymbol{A}$ and $\operatorname{Diag} \boldsymbol{A}$, i.e.,

$$\mathscr{D}_F(\boldsymbol{A}) = \tfrac{1}{2}\|\boldsymbol{A} - \operatorname{Diag} \boldsymbol{A}\|_F^2. \tag{6}$$

This is a general diagonality measure defined for any square matrix not necessarily Hermitian positive-definite. It goes to zero when the sum of squared off-diagonal entries goes to zero independently of the values of the diagonal entries. This measure does not take positive definiteness into account. Furthermore, while this diagonality measure satisfies Property 1, it does not satisfy Property 2.

For a diagonality measure $\mathscr{D}(\cdot)$ to satisfy Property 2, it must be such that $\mathscr{D}(\boldsymbol{A}) = \mathscr{D}(\hat{\boldsymbol{A}})$, where for any positive-definite matrix $\boldsymbol{A}$, the matrix $\hat{\boldsymbol{A}}$ denotes its diagonally scaled matrix given by

$$\hat{\boldsymbol{A}} := (\operatorname{Diag} \boldsymbol{A})^{-1/2} \boldsymbol{A} (\operatorname{Diag} \boldsymbol{A})^{-1/2},$$

i.e., $\hat{a}_{ij} = \dfrac{a_{ij}}{\sqrt{a_{ii}a_{jj}}}$, $1 \le i, j \le n$. We note that $\hat{a}_{ii} = 1$ and $|\hat{a}_{ij}| < 1$ for $1 \le i \ne j \le n$. If $\boldsymbol{A}$ is a positive-definite covariance matrix then $\hat{\boldsymbol{A}}$ is the corresponding correlation matrix. Thus, a properly invariant diagonality measure depends only on the $n(n-1)/2$ independent entries of the hollow (i.e., with vanishing diagonal) symmetric matrix

$$\mathring{\boldsymbol{A}} := (\operatorname{Diag} \boldsymbol{A})^{-1/2} \boldsymbol{A} (\operatorname{Diag} \boldsymbol{A})^{-1/2} - \boldsymbol{I} = \hat{\boldsymbol{A}} - \boldsymbol{I}.$$

3.1. Modified Frobenius measure

Now, using the Frobenius distance between $\hat{\boldsymbol{A}}$ and $\boldsymbol{I}$, we define the modified version of the Frobenius diagonality measure

$$\mathscr{D}_F^m(\boldsymbol{A}) = \tfrac{1}{2}\|(\operatorname{Diag} \boldsymbol{A})^{-1/2} \boldsymbol{A} (\operatorname{Diag} \boldsymbol{A})^{-1/2} - \boldsymbol{I}\|_F^2 = \tfrac{1}{2}\|\hat{\boldsymbol{A}} - \boldsymbol{I}\|_F^2 = \tfrac{1}{2}\|\mathring{\boldsymbol{A}}\|_F^2, \tag{7}$$

which does satisfy Property 2.

3.2. Riemannian measure

The Riemannian measure of diagonality of a Hermitian positive-definite matrix $\boldsymbol{A}$ is defined as half the square of the Riemannian distance (2) between $\boldsymbol{A}$ and $\operatorname{Diag} \boldsymbol{A}$, i.e.,

$$\mathscr{D}_R(\boldsymbol{A}) := \tfrac{1}{2}\|\log((\operatorname{Diag} \boldsymbol{A})^{-1/2} \boldsymbol{A} (\operatorname{Diag} \boldsymbol{A})^{-1/2})\|_F^2 = \tfrac{1}{2}\|\log(\hat{\boldsymbol{A}})\|_F^2. \tag{8}$$

3.3. Kullback-Leibler measure

The *right* Kullback-Leibler measure of diagonality of a symmetric positive-definite matrix $\boldsymbol{A}$ is defined as the Kullback-Leibler divergence [24, 25] between $\boldsymbol{A}$ and $\operatorname{Diag}\boldsymbol{A}$, i.e.,

$$\begin{aligned}\mathscr{D}^r_{KL}(\boldsymbol{A}) &:= \operatorname{tr}(\boldsymbol{A}^{-1}\operatorname{Diag}\boldsymbol{A} - \boldsymbol{I}) - \log\det(\boldsymbol{A}^{-1}\operatorname{Diag}\boldsymbol{A}) \\ &= \operatorname{tr}((\operatorname{Diag}\boldsymbol{A})^{1/2}\boldsymbol{A}^{-1}(\operatorname{Diag}\boldsymbol{A})^{1/2} - \boldsymbol{I}) \\ &\qquad - \log\det((\operatorname{Diag}\boldsymbol{A})^{1/2}\boldsymbol{A}^{-1}(\operatorname{Diag}\boldsymbol{A})^{1/2}) \\ &= \operatorname{tr}\hat{\boldsymbol{A}}^{-1} - n + \log\det\hat{\boldsymbol{A}}. \end{aligned} \tag{9}$$

We note that the fact $\mathscr{D}^r_{KL}(\boldsymbol{A}) \geq 0$, with equality if and only if $\boldsymbol{A}$ is diagonal, is nothing but the fact that, for a positive-definite matrix $\boldsymbol{B}$, $\operatorname{tr}(\boldsymbol{B} - \boldsymbol{I})$ is an upper bound for $\log\det\boldsymbol{B}$, which is attained if and only if $\boldsymbol{B}$ is the identity. This follows immediately from the inequality $\log x \leq x - 1$ which is valid for all $x > 0$ and the equality holds if and only if $x = 1$.

Similarly, the *left* Kullback-Leibler measure of diagonality of a symmetric positive-definite matrix $\boldsymbol{A}$ is defined as the Kullback-Leibler divergence between $\operatorname{Diag}\boldsymbol{A}$ and $\boldsymbol{A}$, i.e.,

$$\begin{aligned}\mathscr{D}^l_{KL}(\boldsymbol{A}) &:= \operatorname{tr}((\operatorname{Diag}\boldsymbol{A})^{-1}\boldsymbol{A} - \boldsymbol{I}) - \log\det((\operatorname{Diag}\boldsymbol{A})^{-1}\boldsymbol{A}) \\ &= \operatorname{tr}((\operatorname{Diag}\boldsymbol{A})^{-1/2}\boldsymbol{A}(\operatorname{Diag}\boldsymbol{A})^{-1/2} - \boldsymbol{I}) \\ &\qquad - \log\det((\operatorname{Diag}\boldsymbol{A})^{-1/2})\boldsymbol{A}(\operatorname{Diag}\boldsymbol{A})^{-1/2}) \\ &= -\log\det\hat{\boldsymbol{A}}. \end{aligned} \tag{10}$$

We note that the fact $\mathscr{D}^l_{KL}(\boldsymbol{A}) \geq 0$, with equality if and only if $\boldsymbol{A}$ is diagonal, is just a form of Hadamard's determinant inequality.

Also, as there are many different ways to symmetrize the Kullback-Leibler divergence, we may define different corresponding diagonality measures. The simplest one is obtained by taking the arithmetic mean of the right and left ones. Hence

$$\mathscr{D}^s_{KL}(\boldsymbol{A}) := \tfrac{1}{2}(\mathscr{D}^r_{KL}(\boldsymbol{A}) + \mathscr{D}^l_{KL}(\boldsymbol{A})) = \tfrac{1}{2}(\operatorname{tr}\hat{\boldsymbol{A}}^{-1} - n). \tag{11}$$

We see that the symmetrized Kullback-Leibler diagonality measure is simply the inverse-barrier function for the set of Hermitian positive-definite matrices, while the left Kullback-Leibler diagonality measure is simply the logarithmic-barrier function for this set.

3.4. Log-det α-measure

For $\alpha \in]-1, 1[$, the log-det α-measure of diagonality of a Hermitian positive-definite matrix $\boldsymbol{A}$ is defined as the log-det α-divergence [22] between $\boldsymbol{A}$ and

Diag $\boldsymbol{A}$, i.e.,

$$\begin{aligned}\mathscr{D}_\alpha(\boldsymbol{A}) &:= \frac{4}{1-\alpha^2}\log\frac{\det(\frac{1-\alpha}{2}\boldsymbol{A}+\frac{1+\alpha}{2}\operatorname{Diag}\boldsymbol{A})}{(\det\boldsymbol{A})^{\frac{1-\alpha}{2}}(\det\operatorname{Diag}\boldsymbol{A})^{\frac{1+\alpha}{2}}}\\ &= \frac{4}{1-\alpha^2}\log\frac{\det(\frac{1-\alpha}{2}\hat{\boldsymbol{A}}+\frac{1+\alpha}{2}\boldsymbol{I})}{(\det\hat{\boldsymbol{A}})^{\frac{1-\alpha}{2}}}. \qquad (12)\end{aligned}$$

The fact that $\mathscr{D}_\alpha(\boldsymbol{A}) \geq 0$, with equality if and only if $\boldsymbol{A}$ is diagonal, is a consequence of Ky Fan's determinant inequality for Hermitian positive-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\det(\lambda\boldsymbol{A}+(1-\lambda)\boldsymbol{B}) \geq (\det\boldsymbol{A})^\lambda(\det\boldsymbol{B})^{1-\lambda}, \quad \forall\lambda \in [0,1].$$

In the special case $\alpha = 0$, we obtain the Bhattacharyya measure of diagonality

$$\mathscr{D}_B(\boldsymbol{A}) := 4\log\frac{\det\frac{1}{2}(\boldsymbol{A}+\operatorname{Diag}\boldsymbol{A})}{\sqrt{\det\boldsymbol{A}\det\operatorname{Diag}\boldsymbol{A}}} = 4\log\frac{\det\frac{1}{2}(\hat{\boldsymbol{A}}+\boldsymbol{I})}{\sqrt{\det\hat{\boldsymbol{A}}}}, \qquad (13)$$

which is half the square of the Bhattacharyya distance [22] between $\boldsymbol{A}$ and Diag $\boldsymbol{A}$.

Using the fact that for an invertible matrix-valued function $\boldsymbol{C}(t)$ we have

$$\frac{d}{dt}\log\det\boldsymbol{C}(t) = \operatorname{tr}\left(\boldsymbol{C}^{-1}(t)\frac{d}{dt}\boldsymbol{C}(t)\right),$$

one can verify that the limit of $\mathscr{D}_\alpha(\boldsymbol{A})$ when α goes to -1 is the right Kullback-Leibler measure and the limit of $\mathscr{D}_\alpha(\boldsymbol{A})$ when α goes to 1 is the left Kullback-Leibler measure.

Remark 2. *All these diagonality measures, except the modified Frobenius one, go to infinity when the matrix $\boldsymbol{A}$ approaches the boundary of the set of Hermitian positive-definite matrices, see Fig. 1 for an illustration of the case of 2×2 matrices.*

3.5. Closed-form expressions for the diagonality measures

3.5.1. Case of 2×2 matrices

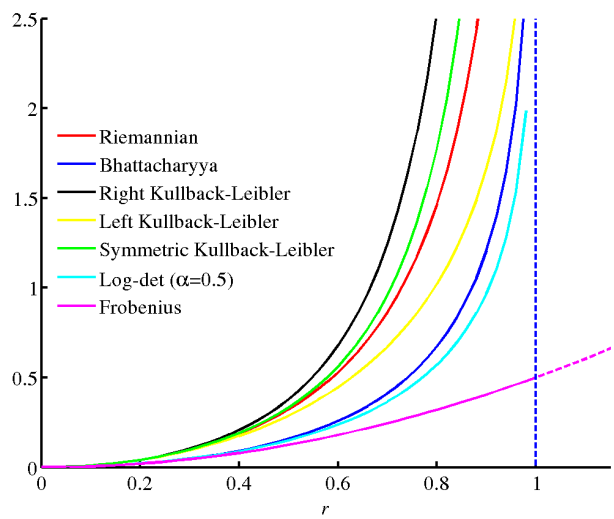
In this case, let $z = \frac{a_{12}}{\sqrt{a_{11}a_{22}}}$ and $r = |z|$. Note that $0 \leq r < 1$ by positive definiteness of $\boldsymbol{A}$. Then,

$$\hat{\boldsymbol{A}} = \begin{bmatrix}1 & z\\ \bar{z} & 1\end{bmatrix}, \quad \mathring{\boldsymbol{A}} = \begin{bmatrix}0 & z\\ \bar{z} & 0\end{bmatrix} \text{ and } r^2 = \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2).$$

Therefore,

$$
\begin{aligned}
\mathscr{D}_F^m(\boldsymbol{A}) &= r^2, \\
\mathscr{D}_R(\boldsymbol{A}) &= \tfrac{1}{2}(\log^2(1+r) + \log^2(1-r)), \\
\mathscr{D}_{KL}^r(\boldsymbol{A}) &= \frac{2r^2}{1-r^2} + \log(1-r^2), \\
\mathscr{D}_{KL}^l(\boldsymbol{A}) &= -\log(1-r^2), \\
\mathscr{D}_{KL}^s(\boldsymbol{A}) &= \frac{r^2}{1-r^2}, \\
\mathscr{D}_\alpha(\boldsymbol{A}) &= \tfrac{2}{1-\alpha^2}\log(1-(\tfrac{1-\alpha}{2}r)^2) - \tfrac{1}{(1+\alpha)}\log(1-r^2), \\
\mathscr{D}_B(\boldsymbol{A}) &= 2\log(1-\tfrac{1}{4}r^2) - \log(1-r^2).
\end{aligned}
$$

Figure 1: Plots of the different diagonality measures for 2-by-2 Hermitian positive-definite matrices. Note that, all diagonality measures except the Frobenius one, are defined for r in $[0, 1[$.



Here we give the power series expansions of these functions about $r = 0$, i.e.,

for almost diagonal matrices:

$$\begin{aligned}
\mathscr{D}_F^m(\boldsymbol{A}) &= r^2, \\
\mathscr{D}_R(\boldsymbol{A}) &= r^2 + \sum_{k=2}^{\infty} \frac{1}{k} \left(\sum_{j=1}^{2k-1} \tfrac{1}{j} \right) r^{2k}, \\
\mathscr{D}_{KL}^r(\boldsymbol{A}) &= r^2 + \sum_{k=2}^{\infty} \tfrac{2k-1}{k} r^{2k}, \\
\mathscr{D}_{KL}^l(\boldsymbol{A}) &= r^2 + \sum_{k=2}^{\infty} \tfrac{1}{k} r^{2k}, \\
\mathscr{D}_{KL}^s(\boldsymbol{A}) &= r^2 + \sum_{k=2}^{\infty} r^{2k}, \\
\mathscr{D}_\alpha(\boldsymbol{A}) &= r^2 + \sum_{k=2}^{\infty} \frac{1}{k} \left(\sum_{j=0}^{k-1} (\tfrac{1-\alpha}{2})^j \right) r^{2k}, \\
\mathscr{D}_B(\boldsymbol{A}) &= r^2 + \sum_{k=2}^{\infty} \frac{2}{k} \left(1 - \tfrac{1}{2^k}\right) r^{2k}.
\end{aligned}$$

3.5.2. Case of 3×3 *matrices*

In this case, let $a = \frac{a_{12}}{\sqrt{a_{11}a_{22}}}$, $b = \frac{a_{31}}{\sqrt{a_{11}a_{33}}}$ and $c = \frac{a_{23}}{\sqrt{a_{22}a_{33}}}$. Then,

$$\hat{\boldsymbol{A}} = \begin{bmatrix} 1 & a & \bar{b} \\ \bar{a} & 1 & c \\ b & \bar{c} & 1 \end{bmatrix}, \quad \mathring{\boldsymbol{A}} = \begin{bmatrix} 0 & a & \bar{b} \\ \bar{a} & 0 & c \\ b & \bar{c} & 0 \end{bmatrix}.$$

Positive definiteness of $\boldsymbol{A}$ implies that $1 - \rho^2 + 2\delta > 0$ and $|\delta| < 1$ where ρ is the non-negative number given by $\rho^2 = |a|^2 + |b|^2 + |c|^2 = \frac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2)$ and $\delta = \operatorname{Re}(abc) = \frac{1}{6}\operatorname{tr}(\mathring{\boldsymbol{A}}^3)$.

Using Cardano's trigonometric solution of a cubic equation [26] the eigenvalues of $\hat{\boldsymbol{A}}$ are $1 + s_i$, $i = 1, \ldots, 3$ where

$$s_1 = \frac{2\rho}{\sqrt{3}} \cos\phi, \quad s_2 = -\frac{2\rho}{\sqrt{3}} \cos\left(\frac{\pi}{3} + \phi\right), \quad s_3 = -\frac{2\rho}{\sqrt{3}} \cos\left(\frac{\pi}{3} - \phi\right),$$

with $\phi = \frac{1}{3}\cos^{-1} \dfrac{3\sqrt{3}\delta}{\rho^3}$.

Then, the diagonality measures are given explicitly by

$$
\begin{aligned}
\mathscr{D}_F^m(\boldsymbol{A}) &= \rho^2,\\
\mathscr{D}_R(\boldsymbol{A}) &= \tfrac{1}{2}(\log^2(1+s_1)+\log^2(1+s_2)+\log^2(1+s_3)),\\
\mathscr{D}_{KL}^r(\boldsymbol{A}) &= \frac{2(\rho^2-3\delta)}{1-\rho^2+2\delta}+\log(1-\rho^2+2\delta),\\
\mathscr{D}_{KL}^l(\boldsymbol{A}) &= -\log(1-\rho^2+2\delta),\\
\mathscr{D}_{KL}^s(\boldsymbol{A}) &= \frac{\rho^2-3\delta}{1-\rho^2+2\delta},\\
\mathscr{D}_\alpha(\boldsymbol{A}) &= \tfrac{2}{1-\alpha^2}\log(1-(\tfrac{1-\alpha}{2}\rho)^2+2(\tfrac{1-\alpha}{2})^3\delta)-\tfrac{1}{1-\alpha}\log(1-\rho^2+2\delta),\\
\mathscr{D}_B(\boldsymbol{A}) &= 2\log(1-\tfrac{1}{4}\rho^2+\tfrac{1}{4}\delta)-\log(1-\rho^2+2\delta).
\end{aligned}
$$

We remark that all measures depend only on the two parameters ρ and δ. Furthermore, it is interesting to note that if δ is equal to zero, i.e., when at least one of the three off-diagonal entries a, b or c is zero, then all these diagonality measures have the same functional expressions as in the 2×2 case (with r replaced by ρ).

3.5.3. The general case of $n\times n$ matrices

In the general case of $n\times n$ matrices, as we shall see, it is possible to have series expansions of the diagonality measures in terms of (only) the $n-1$ invariants, $\operatorname{tr}(\mathring{\boldsymbol{A}}^i)$, $i=2,\dots,n$, of the hollow matrix $\mathring{\boldsymbol{A}}$. Indeed, as the spectral radius of $\mathring{\boldsymbol{A}}$ is less than one, we have $\log(\boldsymbol{I}+\mathring{\boldsymbol{A}})=\sum_{k=1}^\infty \frac{(-1)^{k-1}}{k}\mathring{\boldsymbol{A}}^k$, and $(\boldsymbol{I}+\mathring{\boldsymbol{A}})^{-1}=\sum_{k=0}^\infty(-1)^k\mathring{\boldsymbol{A}}^k$. Therefore, as $\mathring{\boldsymbol{A}}$ is traceless, we obtain

$$
\begin{aligned}
\mathscr{D}_F^m(\boldsymbol{A}) &= \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2),\\
\mathscr{D}_R(\boldsymbol{A}) &= \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2)+\sum_{k=3}^\infty\frac{(-1)^k}{k}\left(\sum_{j=0}^{k-2}\frac{1}{j+1}\right)\operatorname{tr}(\mathring{\boldsymbol{A}}^k),\\
\mathscr{D}_{KL}^r(\boldsymbol{A}) &= \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2)+\sum_{k=3}^\infty(-1)^k\frac{k-1}{k}\operatorname{tr}(\mathring{\boldsymbol{A}}^k),\\
\mathscr{D}_{KL}^l(\boldsymbol{A}) &= \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2)+\sum_{k=3}^\infty\frac{(-1)^k}{k}\operatorname{tr}(\mathring{\boldsymbol{A}}^k),\\
\mathscr{D}_{KL}^s(\boldsymbol{A}) &= \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2)+\sum_{k=3}^\infty\frac{(-1)^k}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^k),\\
\mathscr{D}_\alpha(\boldsymbol{A}) &= \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2)+\sum_{k=3}^\infty\frac{(-1)^k}{k}\left(\sum_{j=0}^{k-2}\left(\frac{1-\alpha}{2}\right)^j\right)\operatorname{tr}(\mathring{\boldsymbol{A}}^k),\\
\mathscr{D}_B(\boldsymbol{A}) &= \tfrac{1}{2}\operatorname{tr}(\mathring{\boldsymbol{A}}^2)+\sum_{k=3}^\infty\frac{2(-1)^k}{k}\left(1-\frac{1}{2^{k-1}}\right)\operatorname{tr}(\mathring{\boldsymbol{A}}^k).
\end{aligned}
$$

Once again, all these measures agree to second order. Now we show that they depend only on the $n-1$ invariants, $T_i := \operatorname{tr}(\mathring{\boldsymbol{A}}^i),\ i = 2, \ldots, n$, of the hollow matrix $\mathring{\boldsymbol{A}}$. Indeed, using the Cayley-Hamilton theorem and since $\mathring{\boldsymbol{A}}$ is a traceless matrix, we have

$$\mathring{\boldsymbol{A}}^n = (-1)^n \sum_{i=2}^{n} b_i \mathring{\boldsymbol{A}}^{n-i},$$

where the coefficients b_i are given by the recursive formula [27]

$$b_i = \frac{1}{i}\left(\sum_{j=2}^{i-2} b_{i-j} T_j + (-1)^n T_i\right), \quad i = 2, \ldots, n.$$

So any power $\mathring{\boldsymbol{A}}^p$ of $\mathring{\boldsymbol{A}}$ with $p > n$ can be expressed in terms of powers $\mathring{\boldsymbol{A}}^q$ of $\mathring{\boldsymbol{A}}$ of order $2 \leq q \leq n$. The result then follows by taking the trace.

4. True diagonality measures

The "true" diagonality measure is defined as the distance (divergence) between a given positive-definite matrix and its closest, with respect to the distance (divergence), diagonal one.

The closest positive-definite diagonal matrix with respect to a distance $d(\cdot,\cdot)$ is defined as

$$\underset{\boldsymbol{X} \in \mathscr{D}_n^{++}}{\arg\min}\, d(\boldsymbol{A}, \boldsymbol{X}).$$

Similarly, the *right* closest positive-definite diagonal matrix with respect to a divergence $D(\cdot,\cdot)$ is defined as

$$\underset{\boldsymbol{X} \in \mathscr{D}_n^{++}}{\arg\min}\, D(\boldsymbol{A}, \boldsymbol{X}),$$

and the *left* closest positive-definite diagonal matrix with respect to a divergence $D(\cdot,\cdot)$ is defined as

$$\underset{\boldsymbol{X} \in \mathscr{D}_n^{++}}{\arg\min}\, D(\boldsymbol{X}, \boldsymbol{A}).$$

The closest, with respect to the Frobenius distance, diagonal matrix to $\boldsymbol{A}$ is $\boldsymbol{D}_F = \operatorname{Diag}(\boldsymbol{A})$, whereas the closest, with respect to the Riemannian distance, is $\boldsymbol{D}_R$ the unique positive-definite diagonal solution to

$$\operatorname{Diag}(\log(\boldsymbol{A}^{-1}\boldsymbol{X})) = \boldsymbol{0}. \tag{14}$$

Equation (14) can be solved numerically by using the steepest descent method for finding the minimum of the convex function $\boldsymbol{X} \mapsto d_{\mathscr{H}_n^{++}}(\boldsymbol{A}, \boldsymbol{X})$ defined on the set of positive-definite diagonal matrices. Starting from an initial guess $\boldsymbol{X}^{(0)}$, we compute the iterates:

$$\boldsymbol{X}^{(k+1)} = \boldsymbol{X}^{(k)} - r\boldsymbol{X}^{(k)} \operatorname{Diag}(\log(\boldsymbol{A}^{-1}\boldsymbol{X}^{(k)})),$$

where r is a conveniently chosen stepsize.

The right closest, with respect to the Kullback-Leibler divergence, diagonal matrix to $\boldsymbol{A}$ is $\boldsymbol{D}^r_{KL} = \mathrm{Diag}(\boldsymbol{A})$, and the left closest, with respect to the Kullback-Leibler divergence, diagonal matrix to $\boldsymbol{A}$ is $\boldsymbol{D}^l_{KL} = (\mathrm{Diag}(\boldsymbol{A}^{-1})^{-1})$. For the symmetrized Kullback-Leibler divergence, the closest diagonal matrix to $\boldsymbol{A}$ is $\boldsymbol{D}^s_{KL} = \mathrm{Diag}(\boldsymbol{A})^{1/2}(\mathrm{Diag}(\boldsymbol{A}^{-1}))^{-1/2}$.

The right closest, with respect to the log-det α-divergence, diagonal matrix to $\boldsymbol{A}$ is $\boldsymbol{D}^\alpha_{LD}$, the unique positive-definite diagonal matrix solution of

$$\mathrm{Diag}\left(\left(\frac{1-\alpha}{2}\boldsymbol{A} + \frac{1+\alpha}{2}\boldsymbol{X}\right)^{-1}\right) = \boldsymbol{X}^{-1}. \tag{15}$$

So, for the Bhattacharyya divergence, the closest diagonal matrix to $\boldsymbol{A}$ is $\boldsymbol{D}_B$ the unique positive-definite diagonal matrix solution of

$$2\,\mathrm{Diag}\left((\boldsymbol{A}+\boldsymbol{X})^{-1}\right) = \boldsymbol{X}^{-1}.$$

Equation (15) can be solved numerically by using the steepest descent method for finding the minimum of the convex function $\boldsymbol{X} \mapsto D^\alpha_{LD}(A, \boldsymbol{X})$ defined on the set of positive-definite diagonal matrices.

We note that $\mathrm{Diag}\,\boldsymbol{A}$ is the closest diagonal matrix to $\boldsymbol{A}$ only for the Frobenius distance and right Kullback-Leibler divergence.

5. Application to the approximate joint diagonalization problem

In Congedo et al. [28] it was shown that the Riemannian metric-based geometric mean [16, 17, 29] of a set of Hermitian positive-definite matrices $\boldsymbol{M}_k$, $k = 1, \ldots, K$, can be approximated numerically by the approximate joint diagonalizer of the set as

$$\boldsymbol{G} \approx \boldsymbol{C}^{-1} \exp\left(\frac{1}{k}\sum_{k=1}^{K} \log \boldsymbol{D}_k\right) \boldsymbol{C}^{-H}, \tag{16}$$

where $\exp\left(\frac{1}{k}\sum_{k=1}^{K} \log \boldsymbol{D}_k\right)$ is the Log-Euclidean mean [30] of the transformed set $\boldsymbol{D}_k = \boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H$, which, thanks to the congruence invariance of the geometric mean, coincides with the transformed geometric mean $\boldsymbol{C}\boldsymbol{G}\boldsymbol{C}^H$ when the transformed matrices all pair-wise commute. More generally, using the same argument, we may approximate all power means of order $p \in [-1, 1]$ (see [31]) by the approximate joint diagonalizer as

$$\boldsymbol{G}_p \approx \boldsymbol{C}^{-1} \left(\frac{1}{k}\sum_{k=1}^{K} \boldsymbol{D}_k^p\right)^{1/p} \boldsymbol{C}^{-H}, \tag{17}$$

where, $\left(\frac{1}{k}\sum_{k=1}^{K} \boldsymbol{D}_k^p\right)^{1/p}$, is the power mean of transformed matrices. Approximation (16) is the limit of (17) when $p \to 0$ and, of course, for $p = 1$ and

$p = -1$ we have the arithmetic mean and harmonic mean, respectively. From a computational point of view the approximate joint diagonalizer approach to compute means may take advantage of the quadratic convergence displayed by some approximate joint diagonalizer algorithms [28]. In [28] it has been shown that the properties of approximation (16) inherit directly from the properties of the cost function used to find the joint diagonalizer $\boldsymbol{C}$; using Pham's cost function the approximation verifies the determinant identity, the joint homogeneity, invariance by congruence transformation, but not the self-duality. An approximate joint diagonalization criterion satisfies the self-duality property whenever if $\boldsymbol{C}$ is an approximate joint diagonalizer of the set $\boldsymbol{M}_1, \ldots, \boldsymbol{M}_K$, then $\boldsymbol{C}^{-H}$ is an approximate joint diagonalizer of the set $\boldsymbol{M}_1^{-1}, \ldots, \boldsymbol{M}_K^{-1}$. Pham's criterion does not satisfy self-duality because it is asymmetric. Using a symmetric criterion the self-duality of the resulting mean would be satisfied as well, thus we would obtain approximations of means with all desirable properties. This extends to all power means. Motivated by this and by recent advances in α-divergence means resulting from the use of the log-barrier function [22, 32, 33], in this article we study approximate joint diagonalization criteria from the perspective of information geometry. While the Riemannian criterion is difficult to handle, we give the gradient and Hessian for a new approximate joint diagonalization cost function based on the α-divergence. For $\alpha = 0$ the resulting approximate joint diagonalization criterion is symmetric, as desired for estimating means of Hermitian positive-definite matrices based on their approximate joint diagonalizer.

The most commonly used cost function in the problem of joint diagonalization of a set of K covariance matrices $\{\boldsymbol{M}_k \in \mathscr{H}_n^{++},\ k = 1, \ldots, K\}$, is Pham's cost function given by [1, 2, 7]

$$\mathcal{J}(\boldsymbol{C}; \{\boldsymbol{M}_k\}) = \sum_{k=1}^{K} \beta_k \left[\log\left(\det(\mathrm{Diag}(\boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H))\right) - \log\left(\det(\boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H)\right)\right], \quad (18)$$

where β_k are positive weights. We remark that the above cost function is nothing but the sum of the weighted left Kullback-Leibler diagonality measures of $\boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H$, $k = 1, \ldots, K$.

We here propose a one-parameter family of cost functions that measure the degree of joint diagonalization of a set of Hermitian positive-definite matrices which is more general than (18) and which includes it as a special case. The family of cost functions $\mathcal{J}_\alpha$, parameterized by $\alpha \in [-1, 1]$, that we propose here is defined as the sum of the log-det α-diagonality measures of the matrices $\boldsymbol{C}\boldsymbol{M}_k\boldsymbol{C}^H$, i.e.,

$$\mathcal{J}^{(\alpha)}(\boldsymbol{C}; \{\boldsymbol{M}_k\}) = \sum_{k=1}^{K} \tilde{\mathcal{J}}^{(\alpha)}_{\boldsymbol{M}_k}(\boldsymbol{C}), \quad (19)$$

where

$$\tilde{\mathcal{J}}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C}) = \mathscr{D}_\alpha(\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H), \quad (20)$$

is the log-det α-diagonality measure of the matrix $\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H$. It should be noted that the log-determinant α-measure of diagonality covers, through the parame-

ter α, all (except the modified Frobenius and Riemannian) diagonality measures discussed in Section 3. We remark that the numerical implementation of an approximate joint diagonalizer based on the Riemannian diagonality measure can be quite complicated. However, the Riemannian diagonality measure can be well approximated by the symmetrized Kullback-Leibler diagonality measure.

The problem of joint diagonalization is then formulated as the following minimization problem

$$\min_{\boldsymbol{C}\in \mathrm{GL}_n(\mathbb{C})} \mathcal{J}^{(\alpha)}(\boldsymbol{C};\{\boldsymbol{M}_k\}),$$

where α is a real parameter in $[-1,1]$ and $\{\boldsymbol{M}_k\}$ is a given set of K Hermitian positive-definite matrices. To numerically solve this minimization problem, we will use a modified Newton method [6]. For this method we need to obtain explicit expressions for the gradient and Hessian of the cost function.

5.1. Gradient and Hessian of the cost function

In order to represent the gradient and the Hessian of the cost function (19) in a compact form, we use the matrix-form representation of the second-order Taylor series expansion as described by Manton in [34] and which we recall here. For a twice differentiable function $\mathcal{J} : \mathcal{M}_n(\mathbb{C}) \to \mathbb{R}$, the second-order Taylor approximation of $\mathcal{J}(\cdot)$ at $\boldsymbol{C}$ reads

$$\begin{aligned}\mathcal{J}(\boldsymbol{C}+t\boldsymbol{Z}) = \mathcal{J}(\boldsymbol{C}) + t\mathfrak{Re}\left\{\mathrm{tr}(\boldsymbol{Z}^H\boldsymbol{G}(\boldsymbol{C}))\right\} + \frac{t^2}{2}\,\mathrm{vec}\,(\boldsymbol{Z})^H\boldsymbol{H}(\boldsymbol{C})\,\mathrm{vec}\,(\boldsymbol{Z}) \\ + \frac{t^2}{2}\mathfrak{Re}\left\{\mathrm{vec}\,(\boldsymbol{Z})^T\boldsymbol{S}(\boldsymbol{C})\,\mathrm{vec}\,(\boldsymbol{Z})\right\} + O(t^3),\end{aligned} \tag{21}$$

where $\boldsymbol{C}$, $\boldsymbol{Z} \in \mathcal{M}_n(\mathbb{C})$, $\boldsymbol{G}(\boldsymbol{C}) \in \mathcal{M}_n(\mathbb{C})$ is the gradient of $\mathcal{J}(\cdot)$ evaluated at $\boldsymbol{C}$, and $\boldsymbol{H}(\boldsymbol{C})$, $\boldsymbol{S}(\boldsymbol{C}) \in \mathcal{M}_{n^2}(\mathbb{C})$ are the Hessians of $\mathcal{J}$ evaluated at $\boldsymbol{C}$. To ensure uniqueness of this representation, $\boldsymbol{H}(\boldsymbol{C})$ is required to be Hermitian and $\boldsymbol{S}(\boldsymbol{C})$ is required to be symmetric, i.e., $\boldsymbol{H}(\boldsymbol{C})^H = \boldsymbol{H}(\boldsymbol{C})$ and $\boldsymbol{S}(\boldsymbol{C})^T = \boldsymbol{S}(\boldsymbol{C})$.

To simplify the derivation of the gradient and Hessian of $\mathcal{J}^{(\alpha)}(\,\cdot\,;\{\boldsymbol{M}_k\})$, it is important to note that $\tilde{\mathcal{J}}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C})$, defined in (20), can be conveniently written as

$$\tilde{\mathcal{J}}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C}) = \frac{4}{1-\alpha^2}\mathcal{J}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C}) - \frac{2}{1+\alpha}\mathcal{J}^{(-1)}_{\boldsymbol{M}}(\boldsymbol{C}) - \frac{2}{1-\alpha}\mathcal{J}^{(1)}_{\boldsymbol{M}}(\boldsymbol{C}), \tag{22}$$

where

$$\mathcal{J}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C}) := \log\det\left(\frac{1-\alpha}{2}\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H + \frac{1+\alpha}{2}\,\mathrm{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H)\right). \tag{23}$$

To obtain the gradient and Hessian of $\mathcal{J}^{(\alpha)}(\boldsymbol{C};\{\boldsymbol{M}_k\})$ it is clear from the expression given in (22) that it suffices to derive the expressions of the gradient and Hessian of $\mathcal{J}^{(\alpha)}_{\boldsymbol{M}}(\cdot)$. In fact, the gradient and Hessian of the function

$\mathcal{J}^{\alpha}(\,\cdot\,;\{\boldsymbol{M}_k\})$ are given by

$$\boldsymbol{G}^{(\alpha)}(\boldsymbol{C};\{\boldsymbol{M}_k\}) = \sum_{k=1}^{K} \widetilde{\boldsymbol{G}}^{(\alpha)}_{\boldsymbol{M}_k}(\boldsymbol{C}),$$

$$\boldsymbol{H}^{(\alpha)}(\boldsymbol{C};\{\boldsymbol{M}_k\}) = \sum_{k=1}^{K} \widetilde{\boldsymbol{H}}^{(\alpha)}_{\boldsymbol{M}_k}(\boldsymbol{C}),$$

$$\boldsymbol{S}^{(\alpha)}(\boldsymbol{C};\{\boldsymbol{M}_k\}) = \sum_{k=1}^{K} \widetilde{\boldsymbol{S}}^{(\alpha)}_{\boldsymbol{M}_k}(\boldsymbol{C}),$$

where $\widetilde{\boldsymbol{G}}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C})$, $\widetilde{\boldsymbol{H}}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C})$ and $\widetilde{\boldsymbol{S}}^{(\alpha)}_{\boldsymbol{M}}(\boldsymbol{C})$ are the gradient and Hessian of $\widetilde{\mathcal{J}}^{(\alpha)}_{\boldsymbol{M}}(\cdot)$ and whose expressions are given in Appendix B.

Once we have derived the gradient and Hessian of our cost function, we now describe a modified Newton. The update at the ℓ-th iteration is given by

$$\boldsymbol{C}^{(\ell+1)} = \boldsymbol{C}^{(\ell)} + \mu_\ell \boldsymbol{W}^{(\ell)},$$

where $\boldsymbol{W}^{(\ell)}$ is the search direction and μ_ℓ is the step size at iteration ℓ. For the Newton method, and since our cost function is not quadratic, we use a modified Newton iteration which includes an Armijo line search. If we are close enough to the minimum the modified Newton update will approach the pure Newton iteration. We note that, as the cost function is not quadratic, the Hessian (whose inverse is needed in the line search) of the cost function can have negative eigenvalues. By adding an appropriate multiple of the identity, the Hessian and hence its inverse can be made positive definite. For implementation issues and further details the reader is referred to [6].

5.2. Numerical experiments

In this section we perform a comparative study of the convergence and performance of the approximate joint diagonalizer algorithm, which is based on the log-det α-diagonality measure and implemented using the Newton method, with two state-of-the-art competing approximate joint diagonalizer algorithms. For the approximate joint diagonalization of the set $\{\boldsymbol{R}_k\}_{k=1}^{K}$ of symmetric positive-definite matrices we use three types of algorithms, the `Jadiag` algorithm [7], which is based on successive transformation operating on a pair of rows and columns, similar to the Jacobi method (but the transformation matrix is not restricted to be orthogonal), the `Uwedge` algorithm [10], which is a low complexity approximate joint diagonalizer algorithm that incorporates nontrivial block-diagonal weight matrices into a weighted least squares approximate joint diagonalization criterion, and `LD-Newton`, which is based on the log-det α-diagonality measure and implemented using the Newton method. It has been shown that the `Jadiag` and the `Uwedge` algorithms have good convergence properties and fast implementation, which gives the comparison process a significant meaning for the evaluation of our method. For the three algorithms, all iterations are started from the identity matrix $\boldsymbol{C}_{(0)} = \boldsymbol{I}_{N\times N}$ as the initial joint diagonalizer of the set $\{\boldsymbol{R}_k\}_{k=1}^{K}$.

In many engineering applications, the matrix condition number of the symmetric positive-definite matrix summarizing the data (observations, recordings, etc.) tends to be positively correlated with the number of sensors. Moreover, the dispersion of the matrices is proportional to the noise level. These properties can be reproduced by the following generating model for the $N \times N$ input data matrices $\{\boldsymbol{R}_k\}_{k=1}^K$ [35]:

$$\boldsymbol{R}_k = \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^T + \nu(\boldsymbol{V}_k\boldsymbol{E}_k\boldsymbol{V}_k^T + \mu\boldsymbol{I}), \tag{24}$$

where, $\boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^T$ is the signal part, $\nu\boldsymbol{V}_k\boldsymbol{E}_k\boldsymbol{V}_k^T$ is the structured noise part and $\nu\mu\boldsymbol{I}$ (we take μ such that $\nu\mu = 10^{-6}$) is the uncorrelated noise part. Here, $\boldsymbol{A}$ is a matrix with elements drawn at random at each simulation from a uniform distribution in $[-1, 1]$ and then normalized so as to have columns with unit norm; $\boldsymbol{D}_k$ are K diagonal matrices with diagonal elements $d_{k,n}$ randomly drawn at each simulation from a squared Gaussian distribution with expectation 2^{-n}, where $n \in \{1, \ldots, N\}$ is the index of the N diagonal elements; $\boldsymbol{V}_k$ are matrices generated as $\boldsymbol{A}$ above; $\boldsymbol{E}_k$ are matrices generated as $\boldsymbol{D}_k$ above and ν is a constant controlling the SNR of the generated matrices through

$$\text{SNR} = \frac{\left[\text{tr}\left(\sum_k \boldsymbol{A}\boldsymbol{D}_k\boldsymbol{A}^T\right)\right]}{\nu\left[\text{tr}\left(\sum_k(\boldsymbol{V}_k\boldsymbol{E}_k\boldsymbol{V}_k^T + \mu\boldsymbol{I})\right)\right]}.$$

In Table 1 we report the CPU time required for each algorithm (implemented in Matlab R2015a running on an Intel Core i5 processor with 4GB RAM on Windows 7 Professional) to converge. For the numerical convergence, we have used the following stopping criterion

$$\frac{1}{N}\|\boldsymbol{C}_{(k)}^{-1}\boldsymbol{C}_{(k-1)} - \boldsymbol{I}\|_F^2 \leq \epsilon,$$

where $\boldsymbol{C}_{(k-1)}$ and $\boldsymbol{C}_{(k)}$ are two successive iterates. The value of ϵ is chosen to be very small (1e15). We note that this criterion does not depend on the size nor on the norm of the matrices.

Since the cost functions of the three algorithms are different, to compare their performance it is meaningful to use the normalized Amari-Moreau performance index [36], which measures the closeness between the estimated joint diagonalizer $\boldsymbol{C}^e$ (the estimated unmixing matrix) and the true unmixing matrix $\boldsymbol{A}^{-1}$, defined by

$$\text{PI}(\boldsymbol{M}) = \frac{1}{2p(p-1)}\left(\sum_{i=1}^p\left(\sum_{j=1}^p \frac{|m_{ij}|}{\max_k|m_{ik}|} - 1\right) + \sum_{j=1}^p\left(\sum_{i=1}^p \frac{|m_{ij}|}{\max_k|m_{kj}|} - 1\right)\right),$$

where m_{ij} are the elements of the matrix $\boldsymbol{M} = \boldsymbol{C}^{eT}\boldsymbol{A}$ and p is the number of sources.

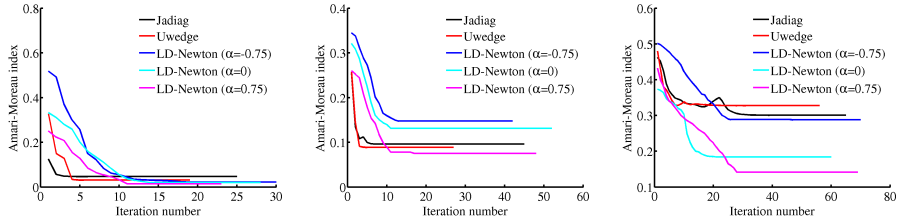
Table 1: The CPU time (in seconds) needed for convergence for each of the `Jadiag`, `Uwedge` and `LD-Newton` (with different values of the parameter α) algorithms in three simulations with different noise levels.

Method	`Jadiag`	`Uwedge`	`LD-Newton`	
SNR $= 100$	0.9360	0.1560	$\alpha = -0.75$	1.6320
			$\alpha = 0$	1.2048
			$\alpha = 0.75$	1.6025
SNR $= 10$	2.4804	1.3276	$\alpha = -0.75$	3.8510
			$\alpha = 0$	3.0351
			$\alpha = 0.75$	3.6124
SNR $= 1$	3.7468	1.9366	$\alpha = -0.75$	5.1097
			$\alpha = 0$	4.1209
			$\alpha = 0.75$	4.7012

The Amari-Moreau performance index is commonly used to estimate the efficiency of (joint) blind source separation algorithms when both the mixing and unmixing matrices are known. This index is invariant under the scaling and permutation ambiguities. Its values are between 0 and 1 and the lower it is the better performance we have.

In Fig. 2 we show an example of convergence plots for the three algorithms, reporting the evolution of the Amari-Moreau index with respect to the iteration number in three simulations. We performed experiments with data matrices are of size $N = 10$ and different SNR taking values in $\{1, 10, 100\}$. For the `LD-Newton` algorithm, the parameter α takes the values $\{-0.75, 0, 0.75\}$.

Figure 2: Plots of the Amari-Moreau index vs the iteration number using `Jadiag`, `Uwedge` and `LD-Newton` (with different values of the parameter α) algorithms. Left: SNR $= 100$, center: SNR $= 10$ and right: SNR $= 1$.



In what follows, we report the Amari-Moreau performance indices using $K = 100$ simulations, by generating different data matrices for every experiment, using the `Jadiag`, `Uwedge` and the `LD-Newton` algorithms. For more complexity, we change the size $N \in \{50, 100\}$ of the matrices to be approximately joint

diagonalized and we increase the noise level up to 0.1. Finally, to get a statistical comparison of the Amari-Moreau indices obtained by the three algorithms, we give the means (averages) for every method through the 100 simulations. Results for different values of $\alpha \in \{-0.75, 0, 0.75\}$ are presented in Fig. 3 and Tab. 2.

Figure 3: Plots of the Amari-Moreau index for $K = 100$ simulations using the `Jadiag`, `Uwedge` and `LD-Newton` (with different values of the parameter α) algorithms for two data sizes ($N = 20$ left, $N = 50$ right).

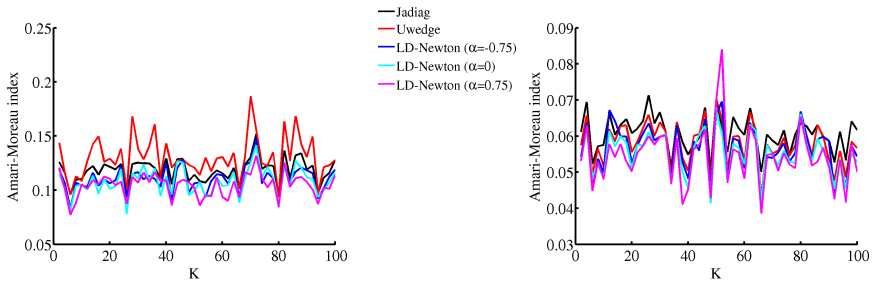


Table 2: The mean (average) for 100 Amari-Moreau indices obtained from 100 different simulations using the `Jadiag`, `Uwedge` and `LD-Newton` (with different values of the parameter α) algorithms for data with $SNR = 0.1$.

Method	`Jadiag`	`Uwedge`	`LD-Newton`	
$N = 20$	0.1162	0.1284	$\alpha = -0.75$	0.1173
			$\alpha = 0$	0.1141
			$\alpha = 0.75$	0.1108
$N = 50$	0.0611	0.0591	$\alpha = -0.75$	0.0574
			$\alpha = 0$	0.0557
			$\alpha = 0.75$	0.0548

The numerical results of the simulations show that these methods are convergent. Even though the `Jadiag` and the `Uwedge` algorithms have the faster convergence in terms of the iterations number and computation time compared with the `LD-Newton`, the latter has proven to have better performance and more accurate results in estimating the approximate joint diagonalizer for the set of data matrices (24). Based on the computations of the Amari-Moreau index, Figs. 2 and 3 show that the `LD-Newton` algorithm has a better robustness for higher noise levels and larger data matrices than the `Jadiag` and the `Uwedge` algorithms. Furthermore, by varying the value of the parameter $\alpha \in \{-0.75, 0, 0.75\}$, through different simulations, the efficiency of the

`LD-Newton` algorithm has demonstrated a slight improvement if we increase the value of α. From the statistical analysis of the three algorithms, illustrated in Tab. 2, the superiority of the `LD-Newton` algorithm, for different value of α, over other algorithms is fairly remarkable.

6. Conclusion

We have introduced several properly-invariant measures for quantifying the closeness of a Hermitian positive-definite matrix to diagonality. These measures are in fact distances or divergences between the given positive-definite matrix and its diagonal part. We gave explicit expressions of these diagonality measures and discussed their invariance properties. We have also defined "true" diagonality measures as distances or divergences between the given positive-definite matrix and its closest diagonal positive-definite matrix. We have used the log-determinant α-measure for the approximate joint diagonalization of a set of Hermitian positive-definite matrices. We described a modified Newton method for the numerical solution of the approximate joint diagonalization problem. The preliminary numerical results show that this method is convergent and has better performance properties than two other algorithms. Convergence analysis and further numerical investigations with real data will be the subject of a future work.

To the best of our knowledge, our algorithm for $\alpha = 0$ is the only geometry-aware approximate joint diagonalizer algorithm using a symmetric function. For instance, the criterion used by Pham's algorithm, which is based on the Kullback-Leibler divergence, is not symmetric. As suggested in [28], the geometric mean computed by an approximate joint diagonalizer approach using our α-divergence would respect all desirable properties.

Appendix A. Useful relations for deriving the gradient and Hessian of the cost function

For $\boldsymbol{Z} \in \mathcal{M}_{m,n}(\mathbb{C})$, the vector $\operatorname{vec}(\boldsymbol{Z}^T) \in \mathbb{C}^{mn}$ is linearly related to the vector $\operatorname{vec}(\boldsymbol{Z}) \in \mathbb{C}^{mn}$. This relation uniquely defines the $mn \times mn$ permutation matrix $\boldsymbol{P}_{m\times n}$ (with entries 0 or 1) through

$$\operatorname{vec}(\boldsymbol{Z}^T) = \boldsymbol{P}_{m\times n} \operatorname{vec}(\boldsymbol{Z}). \tag{A.1}$$

The inverse, $\boldsymbol{P}_{m\times n}^{-1}$, of $\boldsymbol{P}_{m\times n}$ satisfies $\boldsymbol{P}_{m\times n}^{-1} = \boldsymbol{P}_{m\times n}^{T} = \boldsymbol{P}_{n\times m}$. For the special case where $m = n$, the commutation matrix is an involution, i.e., $\boldsymbol{P}_{m\times m}^{2} = \boldsymbol{I}_{m^2}$.

Similarly, for $\boldsymbol{Z} \in \mathcal{M}_m(\mathbb{C})$, the vector $\operatorname{vec}(\operatorname{Diag} \boldsymbol{Z}) \in \mathbb{C}^{m^2}$ is linearly related to the vector $\operatorname{vec}(\boldsymbol{Z}) \in \mathbb{C}^{m^2}$. This relation defines the diagonal projection matrix $\boldsymbol{P}_{\mathrm{Diag}}$ through

$$\operatorname{vec}(\operatorname{Diag} \boldsymbol{Z}) = \boldsymbol{P}_{\mathrm{Diag}} \operatorname{vec}(\boldsymbol{Z}). \tag{A.2}$$

Let $\boldsymbol{Z}$, $\boldsymbol{W} \in \mathcal{M}_{m,n}(\mathbb{C})$ and $\boldsymbol{A}$, $\boldsymbol{B} \in \mathcal{M}_{n,m}(\mathbb{C})$. Then, the following relations hold:

$$\operatorname{tr}(\boldsymbol{Z}\boldsymbol{A}\boldsymbol{W}\boldsymbol{B}) = \operatorname{vec}(\boldsymbol{Z})^T \boldsymbol{P}_{M\times N}^T (\boldsymbol{B}^T \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{W}), \tag{A.3}$$

$$\operatorname{tr}(\boldsymbol{Z}^H\boldsymbol{A}\boldsymbol{W}\boldsymbol{B}) = \operatorname{vec}(\boldsymbol{Z})^H (\boldsymbol{B}^T \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{W}), \tag{A.4}$$

$$\operatorname{tr}(\boldsymbol{Z}^H\boldsymbol{A}\boldsymbol{W}^H\boldsymbol{B}) = \operatorname{vec}(\boldsymbol{Z})^H (\boldsymbol{B}^T \otimes \boldsymbol{A}) \boldsymbol{P}_{m\times n} \operatorname{vec}(\boldsymbol{W})^*. \tag{A.5}$$

For $\boldsymbol{A} \in \mathcal{M}_{p,q}(\mathbb{C})$, $\boldsymbol{B} \in \mathcal{M}_{r,s}(\mathbb{C})$, and $\boldsymbol{Z} \in \mathcal{M}_{q,r}(\mathbb{C})$ we have the relations

$$\operatorname{vec}(\boldsymbol{A}\boldsymbol{Z}\boldsymbol{B}) = (\boldsymbol{B}^T \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{Z}), \tag{A.6}$$

$$\boldsymbol{A} \otimes \boldsymbol{B} = \boldsymbol{P}_{p\times r}(\boldsymbol{B} \otimes \boldsymbol{A})\boldsymbol{P}_{s\times q}. \tag{A.7}$$

Finally, for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ we have

$$\operatorname{tr}(\operatorname{Diag}(\boldsymbol{A})\operatorname{Diag}(\boldsymbol{B})) = \operatorname{tr}(\boldsymbol{A}\operatorname{Diag}(\boldsymbol{B})) = \operatorname{tr}(\operatorname{Diag}(\boldsymbol{A})\boldsymbol{B}). \tag{A.8}$$

Appendix B. Derivation of the gradient and Hessian of the cost function

Before we proceed in the derivation of the gradient and Hessian, we recall the following important result.

Proposition 3 ([6]). *Let $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{M}_n(\mathbb{C})$. Then, for sufficiently small t, we have*

$$\log\left(\det(\boldsymbol{I} + t\boldsymbol{A} + t^2\boldsymbol{B})\right) = t\operatorname{tr}\boldsymbol{A} + t^2\left(\operatorname{tr}\boldsymbol{B} - \tfrac{1}{2}\operatorname{tr}\boldsymbol{A}^2\right) + O(t^3). \tag{B.1}$$

Appendix B.1. Gradient and Hessian of $\mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\cdot)$

From (23) and after expansion we have

$$\begin{aligned}\mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C} + t\boldsymbol{Z}) = \log\det\Bigg(&\frac{1-\alpha}{2}\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H + \frac{1+\alpha}{2}\operatorname{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H)\\ &+ \frac{1-\alpha}{2}(t(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + \boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H) + t^2\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^H)\\ &+ \frac{1+\alpha}{2}\operatorname{Diag}\left(t(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + \boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H) + t^2\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^H\right)\Bigg),\end{aligned}$$

which we can write as

$$\mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C} + t\boldsymbol{Z}) = \mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) + \log\det\left(\boldsymbol{I} + t\boldsymbol{A}_\alpha + t^2\boldsymbol{B}_\alpha\right), \tag{B.2}$$

where, for the sake of simplicity of subsequent development, we have set

$$\boldsymbol{A}_\alpha := (\boldsymbol{Q}_\alpha)^{-1}\left(\tfrac{1-\alpha}{2}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + \boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H) + \tfrac{1+\alpha}{2}\operatorname{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + \boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H)\right),$$
$$\boldsymbol{B}_\alpha := (\boldsymbol{Q}_\alpha)^{-1}\left(\tfrac{1-\alpha}{2}\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^H + \tfrac{1+\alpha}{2}\operatorname{Diag}(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{Z}^H)\right),$$

with $\boldsymbol{Q}_\alpha$ is defined by

$$\boldsymbol{Q}_\alpha := \frac{1-\alpha}{2}\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H + \frac{1+\alpha}{2}\operatorname{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H). \tag{B.3}$$

We note that $\boldsymbol{Q}_{-1} = \boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H$ and $\boldsymbol{Q}_1 = \operatorname{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H)$.

By applying (B.1) to (B.2) we obtain

$$\mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}+t\boldsymbol{Z}) = \mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) + t\operatorname{tr}\boldsymbol{A}_\alpha + t^2\left(\operatorname{tr}\boldsymbol{B}_\alpha - \tfrac{1}{2}\operatorname{tr}(\boldsymbol{A}_\alpha)^2\right) + O(t^3). \tag{B.4}$$

Now, we need to compute the traces of $\boldsymbol{A}_\alpha$, $\boldsymbol{B}_\alpha$ and $(\boldsymbol{A}_\alpha)^2$. Using (A.8) and after some calculations we have

$$\operatorname{tr}\boldsymbol{A}_\alpha = \mathfrak{Re}\left\{\operatorname{tr}\left(\boldsymbol{Z}^H\left((1-\alpha)(\boldsymbol{Q}_\alpha)^{-1} + (1+\alpha)\operatorname{Diag}((\boldsymbol{Q}_\alpha)^{-1})\right)\boldsymbol{C}\boldsymbol{M}\right)\right\}. \tag{B.5}$$

From the expression of $\boldsymbol{B}_\alpha$ and by using (A.4) and (A.8) we obtain

$$\operatorname{tr}(\boldsymbol{B}_\alpha) = \operatorname{vec}(\boldsymbol{Z})^H\boldsymbol{M}^T \otimes \left(\tfrac{1-\alpha}{2}(\boldsymbol{Q}_\alpha)^{-1} + \tfrac{1+\alpha}{2}\operatorname{Diag}((\boldsymbol{Q}_\alpha)^{-1})\right)\operatorname{vec}(\boldsymbol{Z}). \tag{B.6}$$

The trace of $(\boldsymbol{A}_\alpha)^2$ can be conveniently written as

$$\operatorname{tr}(\boldsymbol{A}_\alpha)^2 = \left(\tfrac{1-\alpha}{2}\right)^2 a + \left(\tfrac{1-\alpha^2}{2}\right) b + \left(\tfrac{1+\alpha}{2}\right)^2 c, \tag{B.7}$$

where

$$\begin{aligned}
a &= \operatorname{tr}\left(\boldsymbol{Z}^H(\boldsymbol{Q}_\alpha)^{-1}\left[\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + 2\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H\right](\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M}\right)\\
&\quad + \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\right),\\
b &= \operatorname{tr}\left(\boldsymbol{Z}^H(\boldsymbol{Q}_\alpha)^{-1}\operatorname{Diag}\left(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + \boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H\right)(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M}\right)\\
&\quad + \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\operatorname{Diag}\left(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + \boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H\right)(\boldsymbol{Q}_\alpha)^{-1}\right),\\
c &= \operatorname{tr}\left((\boldsymbol{Q}_\alpha)^{-1}\operatorname{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H)(\boldsymbol{Q}_\alpha)^{-1}\left[\operatorname{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + 2\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H)\right]\right)\\
&\quad + \operatorname{tr}\left((\boldsymbol{Q}_\alpha)^{-1}\operatorname{Diag}(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H)(\boldsymbol{Q}_\alpha)^{-1}\operatorname{Diag}(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H)\right).
\end{aligned}$$

Using (A.3), (A.4), (A.5) and (A.1) we get

$$\begin{aligned}
a &= 2\mathfrak{Re}\left\{\operatorname{vec}(\boldsymbol{Z})^T\boldsymbol{P}_{n\times n}^T\left[(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})^T \otimes (\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})\right]\operatorname{vec}(\boldsymbol{Z})\right\}\\
&\quad + 2\operatorname{vec}(\boldsymbol{Z})^H\left[(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M})^T \otimes (\boldsymbol{Q}_\alpha)^{-1}\right]\operatorname{vec}(\boldsymbol{Z}).
\end{aligned}$$

Similarly, we obtain

$$\begin{aligned}
b &= 2\operatorname{vec}(\boldsymbol{Z})^H\left[\boldsymbol{M}^T\boldsymbol{C}^T(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\right]\boldsymbol{P}_{\mathrm{Diag}}\left[\boldsymbol{C}^*\boldsymbol{M}^T \otimes \boldsymbol{I}_n\right]\operatorname{vec}(\boldsymbol{Z})\\
&\quad + 2\mathfrak{Re}\left\{\operatorname{vec}(\boldsymbol{Z})^T\left[\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1} \otimes (\boldsymbol{Q}_\alpha)^{-T}\right]\boldsymbol{P}_{\mathrm{Diag}}\boldsymbol{P}_{m\times n}\left[\boldsymbol{C}^*\boldsymbol{M}^T \otimes \boldsymbol{I}_n\right]\operatorname{vec}(\boldsymbol{Z})\right.
\end{aligned}$$

Now by use of (A.8), we have

$$\begin{aligned}
c &= \operatorname{tr}\left(\boldsymbol{Z}^H\operatorname{Diag}\left((\boldsymbol{Q}_\alpha)^{-1}\operatorname{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{Z}^H + 2\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H)(\boldsymbol{Q}_\alpha)^{-1}\right)\boldsymbol{C}\boldsymbol{M}\right)\\
&\quad + \operatorname{tr}\left(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H\operatorname{Diag}\left((\boldsymbol{Q}_\alpha)^{-1}\operatorname{Diag}(\boldsymbol{Z}\boldsymbol{M}\boldsymbol{C}^H)(\boldsymbol{Q}_\alpha)^{-1}\right)\right).
\end{aligned}$$

With the help of (A.2), (A.3), (A.4), (A.5) and (A.6) we obtain

$$
\begin{aligned}
c &= \operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}\big(\operatorname{Diag}\big((\boldsymbol{Q}_\alpha)^{-1} \operatorname{Diag}(\boldsymbol{CMZ}^H)(\boldsymbol{Q}_\alpha)^{-1}\big)\big) \\
&\quad + 2\operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}\big(\operatorname{Diag}\big((\boldsymbol{Q}_\alpha)^{-1} \operatorname{Diag}(\boldsymbol{ZMC}^H)(\boldsymbol{Q}_\alpha)^{-1}\big)\big) \\
&\quad + \operatorname{vec}(\boldsymbol{Z})^T \boldsymbol{P}_{n\times n}^T \big[\boldsymbol{I}_n \otimes \boldsymbol{MC}^H\big] \operatorname{vec}\big(\operatorname{Diag}\big((\boldsymbol{Q}_\alpha)^{-1} \operatorname{Diag}(\boldsymbol{ZMC}^H)(\boldsymbol{Q}_\alpha)^{-1}\big)\big) \\
&= \operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \operatorname{vec}\big(\operatorname{Diag}(\boldsymbol{CMZ}^H)\big) \\
&\quad + 2\operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \operatorname{vec}\big(\operatorname{Diag}(\boldsymbol{ZMC}^H)\big) \\
&\quad + \operatorname{vec}(\boldsymbol{Z})^T \boldsymbol{P}_{n\times n}^T \big[\boldsymbol{I}_n \otimes \boldsymbol{MC}^H\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \operatorname{vec}\big(\operatorname{Diag}(\boldsymbol{ZMC}^H)\big).
\end{aligned}
$$

By making use of (A.1) and (A.6) we get

$$
\begin{aligned}
c &= \operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[\boldsymbol{I}_n \otimes \boldsymbol{CM}\big] \boldsymbol{P}_{n\times n} \operatorname{vec}(\boldsymbol{Z})^* \\
&\quad + 2\operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{MC}^H)^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}(\boldsymbol{Z}) \\
&\quad + \operatorname{vec}(\boldsymbol{Z})^T \boldsymbol{P}_{n\times n}^T \big[\boldsymbol{I}_n \otimes \boldsymbol{MC}^H\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{MC}^H)^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}(\boldsymbol{Z}).
\end{aligned}
$$

If we apply (A.7) in the above we find

$$
\begin{aligned}
c &= \operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \boldsymbol{P}_{M\times M} \big[\boldsymbol{CM} \otimes \boldsymbol{I}_n\big] \operatorname{vec}(\boldsymbol{Z})^* \\
&\quad + 2\operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{MC}^H)^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}(\boldsymbol{Z}) \\
&\quad + \operatorname{vec}(\boldsymbol{Z})^T \big[\boldsymbol{MC}^H \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{n\times n} \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{MC}^H)^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}(\boldsymbol{Z}).
\end{aligned}
$$

Finally, since $\boldsymbol{P}_{\mathrm{Diag}} \boldsymbol{P}_{n\times n} = \boldsymbol{P}_{n\times n} \boldsymbol{P}_{\mathrm{Diag}} = \boldsymbol{P}$ we have

$$
\begin{aligned}
c &= 2\operatorname{vec}(\boldsymbol{Z})^H \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{MC}^H)^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}(\boldsymbol{Z}) \\
&\quad + 2\mathfrak{Re}\big\{ \operatorname{vec}(\boldsymbol{Z})^T \big[\boldsymbol{MC}^H \otimes \boldsymbol{I}_n\big] \boldsymbol{P} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{MC}^H)^T \otimes \boldsymbol{I}_n\big] \operatorname{vec}(\boldsymbol{Z}) \big\}.
\end{aligned}
$$

Now, by replacing the expressions of a, b and c in (B.7) and after rearrangement we obtain

$$
\begin{aligned}
\operatorname{tr}(\boldsymbol{A}_\alpha)^2 = 2\operatorname{vec}(\boldsymbol{Z})^H \Bigg[& \left(\frac{1-\alpha}{2}\right)^2 \big[(\boldsymbol{MC}^H(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{CM})^T \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \\
& + \left(\frac{1-\alpha^2}{2}\right) \big[\boldsymbol{M}^T\boldsymbol{C}^T(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{P}_{\mathrm{Diag}} \big[\boldsymbol{C}^*\boldsymbol{M}^T \otimes \boldsymbol{I}_n\big] \\
& + \left(\frac{1+\alpha}{2}\right)^2 \big[(\boldsymbol{CM})^T \otimes \boldsymbol{I}_n\big] \boldsymbol{P}_{\mathrm{Diag}} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{U} \Bigg] \operatorname{vec}(\boldsymbol{Z}) \\
+ 2\mathfrak{Re}\Bigg\{ \operatorname{vec}(\boldsymbol{Z})^T \Bigg[& \left(\frac{1-\alpha}{2}\right)^2 \boldsymbol{P}_{n\times n}^T \big[(\boldsymbol{MC}^H(\boldsymbol{Q}_\alpha)^{-1})^T \otimes (\boldsymbol{MC}^H(\boldsymbol{Q}_\alpha)^{-1})\big] \\
& + \left(\frac{1-\alpha^2}{2}\right) \big[\boldsymbol{MC}^H(\boldsymbol{Q}_\alpha)^{-1} \otimes (\boldsymbol{Q}_\alpha)^{-T}\big] \boldsymbol{P} \big[\boldsymbol{C}^*\boldsymbol{M}^T \otimes \boldsymbol{I}_n\big] \\
& + \left(\frac{1+\alpha}{2}\right)^2 \big[\boldsymbol{MC}^H \otimes \boldsymbol{I}_n\big] \boldsymbol{P} \big[(\boldsymbol{Q}_\alpha)^{-T} \otimes (\boldsymbol{Q}_\alpha)^{-1}\big] \boldsymbol{U} \Bigg] \operatorname{vec}(\boldsymbol{Z}) \Bigg\},
\end{aligned}
\tag{B.8}
$$

where

$$\boldsymbol{U} = \boldsymbol{P}_{\text{Diag}}\left[(\boldsymbol{M}\boldsymbol{C}^H)^T \otimes \boldsymbol{I}_n\right]. \tag{B.9}$$

Inserting (B.5), (B.8) and (B.6) into (B.4) yields

$$\begin{aligned}
\mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}+t\boldsymbol{Z}) = \mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) &+ t\mathfrak{Re}\left\{\text{tr}\left(\boldsymbol{Z}^H\left((1-\alpha)(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M} + (1+\alpha)\,\text{Diag}((\boldsymbol{Q}_\alpha)^{-1})\boldsymbol{C}\boldsymbol{M}\right)\right)\right\} \\
&+ t^2\,\text{vec}\,(\boldsymbol{Z})^H\Bigg(\tfrac{1-\alpha}{2}(\boldsymbol{M}^T\otimes(\boldsymbol{Q}_\alpha)^{-1}) + \tfrac{1+\alpha}{2}(\boldsymbol{M}^T\otimes\text{Diag}((\boldsymbol{Q}_\alpha)^{-1})) \\
&- \left(\tfrac{1-\alpha}{2}\right)^2\left[(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M})^T\otimes(\boldsymbol{Q}_\alpha)^{-1}\right] \\
&- \left(\tfrac{1-\alpha^2}{2}\right)\left[\boldsymbol{M}^T\boldsymbol{C}^T(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\right]\boldsymbol{P}_{\text{Diag}}\left[\boldsymbol{C}^*\boldsymbol{M}^T\otimes\boldsymbol{I}_n\right] \\
&- \left(\tfrac{1+\alpha}{2}\right)^2\left[(\boldsymbol{C}\boldsymbol{M})^T\otimes\boldsymbol{I}_n\right]\boldsymbol{P}_{\text{Diag}}\left[(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\right]\boldsymbol{U}\Bigg)\,\text{vec}\,(\boldsymbol{Z}) \\
&- t^2\mathfrak{Re}\Bigg\{\text{vec}\,(\boldsymbol{Z})^T\Bigg[\left(\tfrac{1-\alpha}{2}\right)^2\boldsymbol{P}_{n\times n}^T\left[(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})^T\otimes(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})\right] \\
&+ \left(\tfrac{1-\alpha^2}{2}\right)\left[\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\otimes(\boldsymbol{Q}_\alpha)^{-T}\right]\boldsymbol{P}\left[\boldsymbol{C}^*\boldsymbol{M}^T\otimes\boldsymbol{I}_n\right] \\
&+ \left(\tfrac{1+\alpha}{2}\right)^2\left[\boldsymbol{M}\boldsymbol{C}^H\otimes\boldsymbol{I}_n\right]\boldsymbol{P}\left[(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\right]\boldsymbol{U}\Bigg]\,\text{vec}\,(\boldsymbol{Z})\Bigg\} + O(t^3).
\end{aligned} \tag{B.10}$$

By comparison between (B.10) and the matrix form of the second-order Taylor expansion (21), we conclude that the gradient $\boldsymbol{G}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C})$ and Hessians $\boldsymbol{H}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C})$ and $\boldsymbol{S}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C})$ of $\mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\cdot)$ evaluated at $\boldsymbol{C}$ are given by

$$\boldsymbol{G}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) = (1-\alpha)(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M} + (1+\alpha)\,\text{Diag}((\boldsymbol{Q}_\alpha)^{-1})\boldsymbol{C}\boldsymbol{M}, \tag{B.11}$$

$$\begin{aligned}
\boldsymbol{H}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) = &(1-\alpha)(\boldsymbol{M}^T\otimes(\boldsymbol{Q}_\alpha)^{-1}) + (1+\alpha)(\boldsymbol{M}^T\otimes\text{Diag}((\boldsymbol{Q}_\alpha)^{-1})) \\
&- \tfrac{(1-\alpha)^2}{2}\left[(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M})^T\otimes(\boldsymbol{Q}_\alpha)^{-1}\right] \\
&- (1-\alpha^2)\left[\boldsymbol{M}^T\boldsymbol{C}^T(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\right]\boldsymbol{P}_{\text{Diag}}\left[\boldsymbol{C}^*\boldsymbol{M}^T\otimes\boldsymbol{I}_n\right] \\
&- \tfrac{(1+\alpha)^2}{2}\left[(\boldsymbol{C}\boldsymbol{M})^T\otimes\boldsymbol{I}_n\right]\boldsymbol{P}_{\text{Diag}}\left[(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\right]\boldsymbol{U},
\end{aligned} \tag{B.12}$$

$$\begin{aligned}
\boldsymbol{S}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) = &-\tfrac{(1-\alpha)^2}{2}\boldsymbol{P}_{n\times n}^T\left[(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})^T\otimes(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})\right] \\
&- (1-\alpha^2)\left[\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\otimes(\boldsymbol{Q}_\alpha)^{-T}\right]\boldsymbol{P}_{\text{Diag}}\boldsymbol{P}_{n\times n}\left[\boldsymbol{C}^*\boldsymbol{M}^T\otimes\boldsymbol{I}_n\right] \\
&- \tfrac{(1+\alpha)^2}{2}\left[\boldsymbol{M}\boldsymbol{C}^H\otimes\boldsymbol{I}_n\right]\boldsymbol{P}_{n\times n}\boldsymbol{P}_{\text{Diag}}\left[(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\right]\boldsymbol{U},
\end{aligned} \tag{B.13}$$

where $\boldsymbol{Q}_\alpha$ is defined in (B.3), $\boldsymbol{U}$ is defined in (B.9) and $\boldsymbol{P} = \boldsymbol{P}_{\text{Diag}}\boldsymbol{P}_{n\times n} = \boldsymbol{P}_{n\times n}\boldsymbol{P}_{\text{Diag}}$.

We note that Joho gave in [6] the expressions of both the gradient and Hessian of the two special cases of $\mathcal{J}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C})$:

$$\begin{aligned}
\mathcal{J}_{\boldsymbol{M}}^{(-1)}(\boldsymbol{C}) &= \log\det(\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H), \\
\mathcal{J}_{\boldsymbol{M}}^{(1)}(\boldsymbol{C}) &= \log\det\text{Diag}(\boldsymbol{C}\boldsymbol{M}\boldsymbol{C}^H).
\end{aligned}$$

Appendix B.2. Gradient and Hessians of $\widetilde{\mathcal{J}}_{\boldsymbol{M}}^{(\alpha)}(\cdot)$

From the definitions (22), (23) and the expressions (B.11), (B.12), (B.13), the gradient and Hessian of $\widetilde{\mathcal{J}}_{\boldsymbol{M}}^{(\alpha)}(\cdot)$ can be expressed as

$$
\begin{aligned}
\widetilde{\boldsymbol{G}}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) &= \tfrac{4}{1-\alpha^2}\left((1-\alpha)\left((\boldsymbol{Q}_\alpha)^{-1}-(\boldsymbol{Q}_{-1})^{-1}\right)+(1+\alpha)\,\mathrm{Diag}((\boldsymbol{Q}_\alpha)^{-1}-(\boldsymbol{Q}_{-1})^{-1})\right)\boldsymbol{C}\boldsymbol{M},\\
\widetilde{\boldsymbol{H}}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) &= \tfrac{4}{1-\alpha^2}\Bigg((1-\alpha)\left(\boldsymbol{M}^T\otimes(\boldsymbol{Q}_\alpha)^{-1}\right)+(1+\alpha)\left(\boldsymbol{M}^T\otimes\mathrm{Diag}((\boldsymbol{Q}_\alpha)^{-1})\right)\\
&\qquad -\tfrac{(1-\alpha)^2}{2}(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\boldsymbol{C}\boldsymbol{M})^T\otimes(\boldsymbol{Q}_\alpha)^{-1}\\
&\qquad -(1-\alpha^2)\big[\boldsymbol{M}^T\boldsymbol{C}^T(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\big]\boldsymbol{P}_{\mathrm{Diag}}\big[\boldsymbol{C}^*\boldsymbol{M}^T\otimes\boldsymbol{I}_n\big]\\
&\qquad -\tfrac{(1+\alpha)^2}{2}\big[(\boldsymbol{C}\boldsymbol{M})^T\otimes\boldsymbol{I}_n\big]\boldsymbol{P}_{\mathrm{Diag}}\big[(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\big]\boldsymbol{U}\Bigg)\\
&\quad -\tfrac{2}{1+\alpha}\left(2\left[\boldsymbol{M}^T\otimes(\boldsymbol{Q}_{-1})^{-1}\right]-2\left[\boldsymbol{M}^T\boldsymbol{C}^T\otimes\boldsymbol{I}_n\right]\left[(\boldsymbol{Q}_{-1})^{-T}\otimes(\boldsymbol{Q}_{-1})^{-1}\right]\left[\boldsymbol{C}^*\boldsymbol{M}^*\otimes\boldsymbol{I}_n\right]\right)\\
&\quad -\tfrac{2}{1-\alpha}\left(2\left[\boldsymbol{M}^T\otimes(\boldsymbol{Q}_1)^{-1}\right]-2\left[\boldsymbol{M}^T\boldsymbol{C}^T\otimes\boldsymbol{I}_n\right]\left[(\boldsymbol{Q}_1)^{-T}\otimes(\boldsymbol{Q}_1)^{-1}\right]\boldsymbol{P}_{\mathrm{Diag}}\left[\boldsymbol{C}^*\boldsymbol{M}^*\otimes\boldsymbol{I}_n\right]\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\widetilde{\boldsymbol{S}}_{\boldsymbol{M}}^{(\alpha)}(\boldsymbol{C}) &= \tfrac{4}{1-\alpha^2}\Bigg(-\tfrac{(1-\alpha)^2}{2}\boldsymbol{P}_{n\times n}^T\big[(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})^T\otimes(\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1})\big]\\
&\qquad -(1-\alpha^2)\big[\boldsymbol{M}\boldsymbol{C}^H(\boldsymbol{Q}_\alpha)^{-1}\otimes(\boldsymbol{Q}_\alpha)^{-T}\big]\boldsymbol{P}\big[\boldsymbol{C}^*\boldsymbol{M}^T\otimes\boldsymbol{I}_n\big]\\
&\qquad -\tfrac{(1+\alpha)^2}{2}\big[\boldsymbol{M}\boldsymbol{C}^H\otimes\boldsymbol{I}_n\big]\boldsymbol{P}\big[(\boldsymbol{Q}_\alpha)^{-T}\otimes(\boldsymbol{Q}_\alpha)^{-1}\big]\boldsymbol{U}\Bigg)\\
&\quad -\tfrac{2}{1+\alpha}\left(-2\left[\boldsymbol{M}\boldsymbol{C}^H\otimes\boldsymbol{I}_n\right]\left[(\boldsymbol{Q}_{-1})^{-1}\otimes(\boldsymbol{Q}_{-1})^{-T}\right]\boldsymbol{P}_{M\times M}\left[\boldsymbol{C}^*\boldsymbol{M}^*\otimes\boldsymbol{I}_n\right]\right)\\
&\quad -\tfrac{2}{1-\alpha}\left(-2\left[\boldsymbol{M}\boldsymbol{C}^H\otimes\boldsymbol{I}_n\right]\left[(\boldsymbol{Q}_1)^{-1}\otimes(\boldsymbol{Q}_1)^{-T}\right]\boldsymbol{P}\left[\boldsymbol{C}^*\boldsymbol{M}^*\otimes\boldsymbol{I}_n\right]\right).
\end{aligned}
$$

Acknowledgment. M.C. and M.M. would like to acknowledge partial support from the European project ERC-2012-AdG-320684-CHESS.

The authors are grateful for the excellent job of the anonymous referee, whose suggestions and remarks helped to improve the early version of this paper.

References

References

[1] B. Flury, W. Gautschi, An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form, SIAM Journal on Scientific and Statistical Computing 7 (1) (1986) 169–184.

[2] B. N. Flury, Common principal components in k groups, Journal of the American Statistical Association 79 (388) (1984) 892–898.

[3] P. Comon, J. Jutten, Handbook of Blind Source Separation, Independent Component Analysis and Applications, Academic Press, Oxford, UK/Burlington, USA, 2010.

[4] J.-F. Cardoso, A. Soulomiac, Blind beamforming for non-Gaussian signals, Proc. Inst. Elect. Eng. F 140 (6) (1993) 362–370.

[5] A.-J. van der Veen, Algebraic methods for deterministic blind beamforming, Proceedings of the IEEE 86 (10) (1998) 1987–2008.

[6] M. Joho, Newton method for joint approximate diagonalization of positive definite Hermitian matrices, SIAM Journal on Matrix Analysis and Applications 30 (3) (2008) 1205–1218.

[7] D. Pham, Joint approximate diagonalization of positive definite Hermitian matrices, SIAM Journal on Matrix Analysis and Applications 22 (4) (2001) 1136–1152.

[8] J. F. Cardoso, A. Souloumiac, Jacobi angles for simultaneous diagonalization, SIAM Journal on Matrix Analysis and Applications 17 (1) (1996) 161–164.

[9] A. Mesloub, K. Abed-Meraim, A. Belouchrani, A new algorithm for complex non-orthogonal joint diagonalization based on shear and Givens rotations, IEEE Transactions on Signal Processing 62 (8) (2014) 1913–1925.

[10] P. Tichavsky, A. Yeredor, Fast approximate joint diagonalization incorporating weight matrices, IEEE Transactions on Signal Processing 57 (3) (2009) 878–891.

[11] A. Yeredor, Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation, IEEE Transactions on Signal Processing 50 (7) (2002) 1545–1553.

[12] A. Yeredor, A. Ziehe, K. R. Müller, Approximate joint diagonalization using a natural gradient approach, in: Independent Component Analysis and Blind Signal Separation, Springer, 2004, pp. 89–96.

[13] A. Ziehe, P. Laskov, K.-R. Müller, A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation, The Journal of Machine Learning Research 5 (2004) 777–800.

[14] R. Bhatia, Pinching, trimming, truncating and averaging of matrices, Amer. Math. Monthly 107 (7) (2000) 602–608.

[15] R. Bhatia, M.-D. Choi, C. Davis, Comparing a matrix to its off-diagonal part, in: H. Dym, S. Goldberg, M. A. Kaashoek, P. Lancaster (Eds.), Operator Theory: Advances and Applications, Vol. 40, Birkhäuser Basel, 1989, pp. 151–164.

[16] M. Moakher, A differential geometric approach to the geometric mean of symmetric positive-definite matrices, SIAM Journal on Matrix Analysis and Applications 26 (3) (2005) 735–747.

[17] R. Bhatia, Positive Definite Matrices, Princeton University Press, New Jersey, 2007.

[18] A. Terras, Harmonic Analysis on Symmetric Spaces and Applications II, Springer-Verlag, New York, 1988.

[19] S.-i. Amari, Differential geometry derived from divergence functions: information geometric approach, in: M. Deza, M. Petitjean, K. Markov (Eds.), Mathematics of Distances and Applications, 1st Edition, ITHEA, Sofia, Bulgaria, 2012, pp. 9–23.

[20] L. M. Bregman, The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics 7 (3) (1967) 200–217.

[21] J. Zhang, Divergence function, duality, and convex analysis, Neural Computation 16 (1) (2004) 159–195.

[22] Z. Chebbi, M. Moakher, Means of Hermitian positive-definite matrices based on the log-determinant α-divergence function, Linear Algebra and its Applications 436 (7) (2012) 1872–1889.

[23] S.-i. Amari, Divergence, optimization and geometry, in: C. S. Leung, M. Lee, J. H. Chan (Eds.), Neural Information Processing: Proceedings of the 16th International Conference (ICONIP 2009), Vol. 5863 of LNCS, Springer-Verlag, Berlin Heidelberg, 2009, pp. 185–193.

[24] M. Moakher, P. G. Batchelor, Symmetric positive-definite matrices: From geometry to applications and visualization, in: J. Weickert, H. Hagen (Eds.), Visualization and Processing of Tensor Fields, Mathematics and Visualization, Springer Berlin Heidelberg, 2006, pp. 285–298.

[25] M. Moakher, Divergence measures and means of symmetric positive-definite matrices, in: D. H. Laidlaw, A. Vilanova (Eds.), New Developments in the Visualization and Processing of Tensor Fields, Mathematics and Visualization, Springer Berlin Heidelberg, 2012, pp. 307–321.

[26] O. K. Smith, Eigenvalues of a symmetric 3×3 matrix, Commun. ACM 4 (4) (1961) 168.

[27] P. Horst, A method for determining the coefficients of a characteristic equation, Ann. Math. Statist. 6 (2) (1935) 83–84.

[28] M. Congedo, B. Afsari, A. Barachant, M. Moakher, Approximate joint diagonalization and geometric mean of symmetric positive definite matrices, PLoS ONE 10 (2015) e0121423.

[29] R. Bhatia, J. Holbrook, Riemannian geometry and matrix geometric means, Linear Algebra Appl. 413 (2006) 594–618.

[30] V. Arsigny, P. Fillard, X. Pennec, N. Ayache, Geometric means in a novel vector space structure on symmetric positivedefinite matrices, SIAM Journal on Matrix Analysis and Applications 29 (1) (2007) 328–347.

[31] Y. Lim, M. Pálfia, Matrix power means and the Karcher mean, Journal of Functional Analysis 262 (4) (2012) 1498–1514.

[32] M. Pálfia, Operator means of probability measures and generalized karcher equations, Advances in Mathematics 289 (2016) 951–1007.

[33] S. Sra, Positive definite matrices and the S-divergence, Proc. Amer. Math. Soc. (2015) 12953.

[34] J. H. Manton, Optimization algorithms exploiting unitary constraints, IEEE Transactions on Signal Processing 50 (3) (2002) 635–650.

[35] M. Congedo, R. Phlypo, A. Barachant, A fixed-point algorithm for estimating power means of positive definite matrices, in: Proceedings of the EUSIPCO Conference, Budapest, Hungary, 2016.

[36] O. Macchi, E. Moreau, Self-adaptive source separation by direct and recursive networks, in: Proceedings of the International Conference on Digital Signal Processing, 1993, pp. 1154–1159.