Blind separation with angular criteria

Romain Lloria, Florent Bouchard, Sylvain Chevallier, Frédéric Pascal

Blind separation with angular criteria

Romain Lloria, Florent Bouchard, Sylvain Chevallier, Frédéric Pascal

Contents

1 Blind source separation

We are interested in the blind source separation (BSS) problem, which was initiated in [16, 32]; see [17] for a full review of the problem. In this report, we consider the instantaneous linear mixing model [17] $\mathcal{X}(t) = A\mathcal{S}(t)$, where $\mathcal{S}(t) \in \mathbb{R}^n$ and $\mathcal{X}(t) \in \mathbb{R}^n$ are the source and observed signals, and $A \in \mathrm{GL}_n(\mathbb{R})$ (general linear group) is the so-called mixing matrix. Both signals are defined on a set $\mathcal{S}$, $\mathcal{X} : \Omega \subset \mathbb{R} \longrightarrow \mathbb{R}^n$ with depends on the type of signals. For audio signals, it is a time interval $\Omega \subset \mathbb{R}$. For MRI data it is a set of spatial coordinates $\Omega \subset \mathbb{R}^3$. In the case of image analysis it is a set of frequencies. Given some observations $\mathcal{X}(t)$, the goal is to find estimates, again denoted A and $\mathcal{S}(t)$, of the true mixing matrix and sources. We do it only assuming statistical independence of the sources. To do so, one usually only estimates an unmixing matrix $B \in \mathrm{GL}_n(\mathbb{R})$. We then get $A = B^{-1}$ and $\mathcal{S}(t) = B\mathcal{X}(t)$. Notice that the BSS problem has two inherent ambiguities: permutation and diagonal scaling. Indeed, for any non-singular diagonal matrix $D \in D_n(\mathbb{R})^*$ (non-singular diagonal matrices group) and permutation matrix $P \in \mathcal{P}_n(\mathbb{R})$ (permutation matrices group), B and PDB are equivalent solutions.

As explained, only the statistical independence of sources is assumed to retrieve the sources and mixing process from the observations. To identify them, one needs to exploit some sort of statistical diversity from the observations. Two different paradigms exist:

- independent and identically distributed (iid) non-Gaussian sources: when assumed iid, sources need to be non-Gaussian to have enough statistical diversity and identify them. In such a case, higher-order statistics (cumulants, mutual information, etc.) are exploited [16, 11, 40, 17].
- non-iid Gaussian sources: since only their second-order moment is non-zero, they must be non-iid to have enough statistical diversity for Gaussian sources. Usually, one then exploits non-stationarity [42] (temporal diversity) or coloration [18] (frequency diversity). This way, one can construct several covariance matrices from the observations.

In this thesis, we are interested in the second approach (non-iid Gaussian sources). To achieve the separation, we construct K iid centered Gaussian random variables $\{\mathcal{X}_k\}_{k=1}^K$ induced by K iid centered Gaussian random variables $\{\mathcal{S}_k\}_{k=1}^K$[1]. Since they are centered, Gaussian vectors $\{\mathcal{S}_k\}_{k=1}^K$ and $\{\mathcal{X}_k\}_{k=1}^K$ are solely characterized by their covariance matrices $\{\mathrm{Cov}\mathcal{S}_k\}_{k=1}^K$ and $\{\mathrm{Cov}\mathcal{X}_k\}_{k=1}^K$, where, $\forall k \in [\![1, K]\!]$, $\mathrm{Cov}\mathcal{X}_k = A\mathrm{Cov}\mathcal{S}_k A^T$ by linearity. Furthermore, the statistical independence of the sources implies that their covariance matrices $\{\mathrm{Cov}\mathcal{S}_k\}_{k=1}^K$ are diagonal. In practice, for each random variable $\mathcal{X}_k$, one has N iid samples $\{\mathcal{X}_k(t_i)\}_{i=1}^N$ and the covariance matrix $\mathrm{Cov}\mathcal{X}_k$ can be estimated with the sample covariance matrix (SCM) estimator $\mathrm{Cov}\mathcal{X}_k = \frac{1}{N}\sum_{i=1}^N \mathcal{X}_k(t_i)\mathcal{X}_k(t_i)^T$. From these estimates $\{\mathrm{Cov}\mathcal{X}_k\}_{k=1}^K$, one aims to find estimates $\{\mathrm{Cov}\mathcal{S}_k\}_{k=1}^K$ of the covariance matrices of the sources. To do so, the goal is to find $B \in \mathrm{GL}_n(\mathbb{R})$ such that, $\forall k \in [\![1, K]\!]$, $\mathrm{Cov}\mathcal{S}_k = B\mathrm{Cov}\mathcal{X}_k B^T$ are as diagonal as possible. Thus, in our case, the BSS problem boils down to solving the approximate joint diagonalization problem of the set of covariance matrices $\{\mathrm{Cov}\mathcal{X}_k\}_{k=1}^K$. Now, we want to minimize simultaneously the diagonality's degree of the $x\mathrm{Cov}\mathcal{X}_k x^T$. We define this simultaneous diagonality criterion by doing the sum

$$\phi(x) = \sum_{k=1}^{K} \mathcal{D}(x\mathrm{Cov}\mathcal{X}_k x^T) \tag{1}$$

where $\mathcal{D}$ is a way to measure the degree of diagonality of $x\mathrm{Cov}\mathcal{X}_k x^T$. We do it implementing the gradient algorithm. In other words, we want to find the minimum of a simultaneous diagonality criterion

$$B = \underset{x \in \mathrm{GL}_n(\mathbb{R})}{\mathrm{argmin}}\ \phi(x) \tag{2}$$

Review The two first historical diagonality measures were the least-square criterion and the Kullback-Leibler divergence. The least-square criterion minimizes the sum of the squared non-diagonal elements [11, 7]. The Kullback-Leibler divergence comes from the non-iid Gaussian statistical model. It measures the similarity between $x\mathrm{Cov}\mathcal{X}_k x^T$ and its diagonal part [41]. Then, both were generalized using divergences to measure a similarity $d(S, D_n^{++}(\mathbb{R}))$ between $S \in S_n^{++}(\mathbb{R})$ and $D_n^{++}(\mathbb{R})$ (set of positive definite diagonal matrices). More precisely, a divergence measures the dissimilarity between two points. So we consider the closest diagonal matrix from S, denoted Λ_S, with respect to d. If Λ_S exists, we obtain a well defined measure of diagonality $\mathcal{D}$ with $\mathcal{D}(S) = d(S, \Lambda_S)$. We can see

[1] To obtain these, one can, for instance, exploit non-stationarity: $\{\mathcal{X}_k\}_{k=1}^K$ then correspond to K different windows of the data. For coloration, the random variables correspond to K different frequencies (for example obtained through Fourier transform). See, e.g., [42, 18] for more details.

that it cancels if and only if S is diagonal. We strongly recommend reading [5] on an excellent presentation of this object. For example, the square of a distance is a divergence. Likewise, we can measure the degree of diagonality of $S \in S_n^{++}(\mathbb{R})$ using its Riemannian distance (or the square to have a divergence) with the closest diagonal matrix [10]. Many interesting questions lie in the links between distances and divergences [46].

Concerning optimization methods, many iterative algorithms have been developed for the two historical criteria. The first looked for x in $O_n(\mathbb{R})$ [11]. Then, the community looked for $x \in \mathrm{GL}_n(\mathbb{R})$ deleting previous whitening errors. For example, we found methods using Jacobi algorithm [11, 12, 23], Lagrange multipliers [19, 49], Newton [29, 31], quasi-Newton [43] or fixed point theorem [48]. Finally, [28, 44, 47] propose Riemannian optimization methods in the orthogonal case. Furthermore, constraints are added to avoid aberrant solutions (like 0). For example, x could be searched in $\mathrm{GL}_n(\mathbb{R})$ such that the norms of its columns are equal to 1. Several strategies have been proposed to add these constraints. For example, [11, 12, 23] resorts to transformations that preserve the constraints. Some studies impose constraints at the end of each update [43, 48]. Others compensate by adding an error term to the cost function [30, 31]. Finally, [2] proposes to optimize on constraint-submanifolds in $\mathrm{GL}_n(\mathbb{R})$, giving a general optimization framework [9, 10, 37].

Thesis objectives Several objectives are set for this thesis work.

- Develop new diagonality criteria based on Riemannian angles. The idea is to measure the proximity of $x\mathrm{Cov}\mathcal{X}_k x^T$ with $D_n^{++}(\mathbb{R})$ by measuring the angle (in the identity) between the tangent space and the geodesic joining I_n and $x\mathrm{Cov}\mathcal{X}_k x^T$. We will study their properties and will be interested in the geometric information they contain. In searching for diagonality criteria's unification, looking for links with previous criteria will be interesting. Then, we evaluate the performances by comparison with the state of the art. Finally, it would be interesting to combine our new criteria and use the strong points of each of them.

- The question of robustness will also be studied. For example, estimating covariance matrices with robust estimators (like maximum likelihood or elliptic distribution or M estimators [36, 50]), rather than SCM, could give better performance when containing noise and aberrant values.

- Finally, we will mix the two approaches.

Work done so far First, a long period was devoted to getting to grips with coding tools, the discovery of computer simulation, the source separation model, and its context. Then, we have defined new diagonality criteria based on Riemannian angles and robust statistics. Two contributions were made.

- For the first contribution, we use inherent ambiguities to propose a novel Riemannian framework, simplifying the previous one. Rather than optimizing directly on $\mathrm{GL}_n(\mathbb{R})$ or using polar decomposition, we propose to use LU decomposition. Unlike the open set $\mathrm{GL}_n(\mathbb{R})$, using invertible triangular matrices gives considerable numerical advantages. Moreover, these spaces are closed, which makes it possible to ensure that the points of convergence are also there. Moreover, because of the inherent ambiguities, aberrant solutions may appear. The LU decomposition will also allow us to reduce this ambiguity.

- Our second contribution lies in the following section. Using Riemannian angles, we define new diagonality criteria for a general complete metric on $S_n^{++}(\mathbb{R})$. We study their properties and calculate their gradients to implement the previous framework.

  Then, we will note that angles only give partial information and will correct this defect by adding a distance term. So, we obtain a parametrized criteria family. We obtain already known divergences for specific values of this parameter (with Euclidean and affine-invariant metrics). Conversely, other known divergences are written as angular criteria. These results are part of the previously presented unification research.

  Then, we will evaluate performances using Euclidean and affine-invariant metrics. For this, we will use the previous framework. We will see that the Euclidean angular criteria do better for low noise than the Kullbak-Leibler divergence. Then, we study the influence of the previous parameter on performance.

2 Novel optimization framework with LU decomposition

We are interested in the approximated joint diagonalization problem rewritten as an optimization problem (2). To simplify this optimization problem, we use the LU decomposition. In this section, we develop the corresponding optimization framework.

As $x \in \mathrm{GL}_n(\mathbb{R})$, it admits a PLU decomposition [25], we have $x = Px_Lx_U$ where P is a permutation matrix, $x_L \in \mathbb{L}^\times$ (invertible lower triangular matrices) and $x_U \in \mathbb{U}^1$ (upper triangular matrices with 1 on the diagonal). Thanks to the permutation ambiguity of the BSS problem, we bring back to optimization on $\mathbb{L}^\times \times \mathbb{U}^1$ and still access all possible solutions. So, the problem can be rewritten in this space, and we want to implement the gradient algorithm.

In section 2.1, we present the corresponding subspaces and calculate the corresponding gradient of ϕ. In section 2.2, we take care of diagonal scaling ambiguity with gives aberrant solutions (for example, the Frobenius divergence admits 0 as an aberrant solution). To do this, we add a constraint with 1 on the diagonal on the two triangular matrices. It is equivalent to adding a constraint on x_L from the previous optimization problem. We present the subspaces and calculate the corresponding gradient of ϕ. In section 2.3, we present the Kullback-Leibler divergence. We will use it as a reference diagonality criterion to evaluate the performance of the LU algorithm. Section 2.4 compares this framework with the equivalent framework on $\mathrm{GL}_n(\mathbb{R})$.

2.1 LU decomposition: the riemLU algorithm

As explained before, we look for x of the form $x = x_Lx_U$, where $x_L \in \mathbb{L}^\times$. $\mathbb{L}^\times \times \mathbb{U}^1$ is a product affine space, whose tangent space at $(x_L, x_U) \in \mathbb{L}^\times \times \mathbb{U}^1$ is $\mathbb{L} \times \mathbb{U}^0$, where $\mathbb{L}$ and $\mathbb{U}^0$ are the sets of lower and strict upper triangular matrices. We endow this affine space with the product Euclidean metric $g_{F\times F} = g_F + g_F$ from $M_n(\mathbb{R}) \times M_n(\mathbb{R})$ ($M_n(\mathbb{R})$ is the general matrices set). We rewrite the problem on $\mathbb{L}^\times \times \mathbb{U}^1$ with the diffeomorphism (on its image)

$$p : \left\{ \begin{array}{ccc} \mathbb{L}^\times \times \mathbb{U}^1 & \longrightarrow & \mathrm{GL}_n(\mathbb{R}) \\ (x_L, x_U) & \longrightarrow & x_Lx_U \end{array} \right.$$

Thus we optimize $\overline{\phi} = \phi \circ p : \mathbb{L}^\times \times \mathbb{U}^1 \longrightarrow \mathbb{R}$. By composition, this map is differentiable, and a direct calculation gives the gradient with respect to the Euclidean gradient $\mathrm{grad}\,\mathcal{D}$ of the chosen diagonality criteria $\mathcal{D}$. We get

$$\begin{aligned} \mathrm{grad}^{\mathbb{L}^\times\times\mathbb{U}^1}\overline{\phi}(x_L, x_U) =& 2\sum_{k=1}^{K} \mathrm{trig} \left[\begin{array}{c} \mathrm{grad}\,\mathcal{D}(x_Lx_U\mathrm{Cov}\mathcal{X}_kx_U^Tx_L^T)x_Lx_U\mathrm{Cov}\mathcal{X}_kx_U^T \\ x_L^T\mathrm{grad}\,\mathcal{D}(x_Lx_U\mathrm{Cov}\mathcal{X}_kx_U^Tx_L^T)x_Lx_U\mathrm{Cov}\mathcal{X}_k \end{array} \right] \\ &+ \left[\begin{array}{c} \mathrm{diag}\,\left\{\mathrm{grad}\,\mathcal{D}(x_Lx_U\mathrm{Cov}\mathcal{X}_kx_U^Tx_L^T)x_Lx_U\mathrm{Cov}\mathcal{X}_kx_U^T\right\} \\ 0 \end{array} \right] \end{aligned} \tag{3}$$

$\mathrm{diag}\,M$ is the diagonal par of M and $\mathrm{trig}\,(M, M') = (M_-, M'_+)$ with M_- the strict lower part of M and M'_+ the strict upper part of M'. Now, we can implement the previous gradient algorithm, giving the called riemLU algorithm.

2.2 Unitary LU decomposition: the riemUnit algorithm

As explained in the introduction, we now $x = x_Lx_U$, where $x_L \in \mathbb{L}^1$ (lower triangular matrices with 1 on the diagonal) and $x_U \in \mathbb{U}^1$ (upper triangular matrices with 1 on the diagonal). $\mathbb{L}^1 \times \mathbb{U}^1$ is an affine product space whose tangent spaces are $\mathbb{L}^0 \times \mathbb{U}^0$ (strict trigonal matrices). We equip it with the same product, Euclidean metric. Similar calculation gives

$$\mathrm{grad}^{\mathbb{L}^1\times\mathbb{U}^1}\overline{\phi}(x_L, x_U) = 2\sum_{k=1}^{K} \mathrm{trig} \left[\begin{array}{c} \mathrm{grad}\,\mathcal{D}(x_Lx_U\mathrm{Cov}\mathcal{X}_kx_U^Tx_L^T)x_Lx_U\mathrm{Cov}\mathcal{X}_k\,x_U^T \\ x_L^T\mathrm{grad}\,\mathcal{D}(x_Lx_U\mathrm{Cov}\mathcal{X}_kx_U^Tx_L^T)x_Lx_U\mathrm{Cov}\mathcal{X}_k \end{array} \right] \tag{4}$$

2.3 Comparaison algorithm: Kullback-Leibler divergence

We use the left Kullback-Leibler divergence as a diagonality criterion to evaluate our algorithm. Following [17, 42], we present it quickly. This divergence is written $\mathcal{D}_{KL}(\boldsymbol{S}) = \ln\det(\mathrm{diag}\,\boldsymbol{S}) - \ln\det\,\boldsymbol{S}$ for $\boldsymbol{S} \in S_n^{++}(\mathbb{R})$. We obtain the cost function by injecting it into (1). By composition with the determinant, this function is differentiable, and the fact that the partial differential at the closest diagonal matrix is zero gives directly $\mathrm{grad}\,\mathcal{D}_{KLl}(S) = \mathrm{diag}^{-1}S - S^{-1}$ We inject this equality into (3) and (4) to obtain the corresponding gradient.

2.4 Numerical experiment

Now we compare our riemLU (2.1) and riemUnit (2.2) performances with the Euclidean space one (riemEucl) and the qndiag algorithm of [1] based on the maximum likelihood estimator [42]. In the same way as [17, 42], we propose to do it with the previous left Kullback-Leibler divergence presented previously.

Data simulation We generate the data as follows. Mixing matrix $A = \boldsymbol{O} \cdot \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\boldsymbol{T}}$; $\boldsymbol{O}, \boldsymbol{U} \sim \mathcal{U}[O_n(\mathbb{R})]$, $\boldsymbol{D} \sim \mathcal{U}\left(\left[\frac{1}{10}, 10\right]^n\right)$ and $\operatorname{Cond} A = 5$. Diagonals initial covariance matrices (centered reduced) $\operatorname{Cov}\mathcal{S}_k \sim \chi^2(4)$. Observed covariance matrix $\operatorname{Cov}\mathcal{X}_k = A\operatorname{Cov}\mathcal{S}_k A^T$. As explained in the introduction, in each Monte-Carlo's loop, we generate N random vectors $X_k \backsim \mathcal{N}(0_n, \operatorname{Cov}\mathcal{X}_k)$ and determine an approximation by SCM of the $\operatorname{Cov}\mathcal{X}_k$. According to whitening, we implement our four algorithms to minimize the measure of simultaneous diagonality ϕ_{KL} and obtain the approximation of $\widehat{A}^{-1}$ up to permutation matrix and diagonal scaling.

Calculation of errors We compare with the Riemannian error (the sum of the squares of the Riemannian distances in $S_n^{++}(\mathbb{R})$ between the $x\operatorname{Cov}\mathcal{X}_k\, x^T$ and their approximate) and the Amari criterion [35] (which measures the similarity between x and A^{-1} up to permutation and diagonal scaling). With $\boldsymbol{M} = xA$ Amari criterion is

$$\frac{1}{2n(n-1)} \sum_{i=1}^{n} \left(\frac{\sum_{j=1}^{n} |\boldsymbol{M}_{ij}|}{\operatorname{Max}_{1 \leqslant j \leqslant n} |\boldsymbol{M}_{ij}|} + \frac{\sum_{i=1}^{n} |\boldsymbol{M}_{ji}|}{\operatorname{Max}_{1 \leqslant i \leqslant n} |\boldsymbol{M}_{ji}|} - 2 \right)$$

Parameters We draw mean and median errors after several Monte-Carlo loops, for the K observations. We do it with $n = 8$, 100 random vectors, $N = 50$ Monte-Carlo loops, and the graphs are plotted with $K = 10; 30; 50; 75$, obtaining (1).

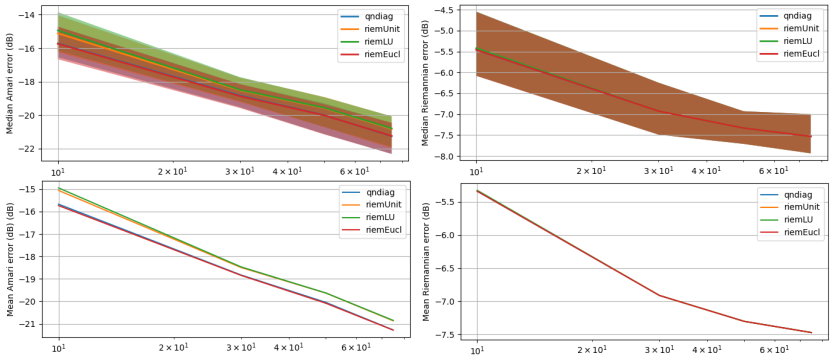

Figure 1: Mean and median error

Conclusions We remark that

- riemUnit, riemLU and riemEucl give the same Riemannian error as the reference qndiag
- riemUnit and riemLU algorithms give slightly higher Amari errors than qndiag
- as expected, riemEucl gives the same Amari error as the qndiag algorithm. It gives an additional argument against riemUnit and riemLU

With this type of simulation, riemUnit and riemLU are slightly less efficient than qndiag or riemEucl for the Amari criterion. Moreover, the performance looks comparable to the Riemannian distance error.

3 Novel diagonality criteria based on angles

So far, our approximate joint diagonalization problem has been treated using similarity measures between S and its closest diagonal matrix (for the chosen measure). This section proposes an original direction by defining a similarity measure based on angles [38, 24]. To our knowledge, this is the first time that Riemannian angles are used in this context.

In section 3.1, we quickly explain how we use it to construct an angular criterion, and then we calculate it. We will see that its gradient diverges if S is diagonal, so consider alternative criteria to achieve our goals. Section 3.2 highlights that angular criteria only give partial information. We correct this by ponderating with a distance term. With a parameter β, we control its influence, obtaining a family of parameterized criteria. Section 3.3 highlights that these criteria allow us to unify several already known divergences. Section 3.4 compares our criteria with the previously introduced riemEucl algorithm. We will also try to determine the influence of β on these parameters.

3.1 The angular criteria

Here, the manifold is $S_n^{++}(\mathbb{R})$ with a complete Riemannian metric g (Euclidean or affine-invariant).

3.1.1 Construction of the angular criterion

Our criteria will be the measure of the Riemannian angle at I_n between S and $\mathbb{T}_{\mathrm{I}_n} D_n^{++}(\mathbb{R}) = D_n(\mathbb{R})$ (diagonal matrices set). For this, we choose a representative $\Lambda \in D_n^{++}(\mathbb{R})$, and measure the angle in I_n between the minimizing geodesics, which join I_n with respectively S and Λ. As a representative, we propose to choose the nearest diagonal matrix Λ_S of S for the Riemannian angle measure. It will allow us to have the central property: S is diagonal iff $\angle^g(S) = 0$. We know that such a Λ_S exists thanks to Cauchy-Schwarz inequalitie. Calculating geodesics (B.2), the definition of the Riemannian angle gives our angular criterion.

Définition 3.1 (Angular criterion)**.** The angular criterion on $S \in S_n^{++}(\mathbb{R})$ is

$$\angle^g(S) = \arccos \frac{\langle \ln^g_{\mathrm{I}_n} S, \ln^g_{\mathrm{I}_n} \Lambda_S \rangle_{\mathrm{I}_n}}{\| \ln^g_{\mathrm{I}_n} S\|_{\mathrm{I}_n} \| \ln^g_{\mathrm{I}_n} \Lambda_S\|_{\mathrm{I}_n}}$$

where $\ln^g_{\mathrm{I}_n} = \left(\exp^g_{\mathrm{I}_n}\right)^{-1}$ well-defined by completeness.

By abuse, we will similarly denote with $\angle^g$ the angular criterion and $\angle^g(S_1, S_2)$ the measure of the angle (in the identity) between the geodesic joining I_n to S_1 and the geodesic joining I_n to S_2. Of course, the angular criterion is defined for $S \neq \mathrm{I}_n$; otherwise S is already diagonal.

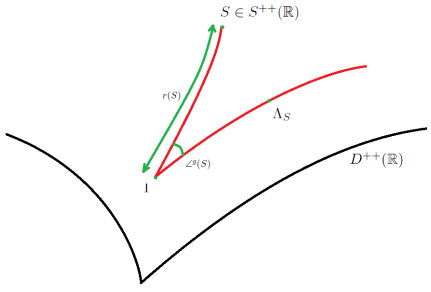

Figure 2: angle

Now, it is necessary to prove that this new criterion has the desired properties (D).

Propriétés 3.2. *The Angular criterion $\angle^g$ verifies the following properties.*
(1) $\angle^g \in [0, \pi/2]$.
(2) $\angle^g$ is at least twice differentiable.
(3) $\angle^g = 0$ iff S is diagonal.

Now, we have clearly defined our angular criterion, and we can explicitly calculate it (E). The closest diagonal matrices (for $\angle^g$) are all the points of the geodesic between I_n and $\Lambda_S = \exp^g_{\mathrm{I}_n} \operatorname{diag} \ln^g_{\mathrm{I}_n} S$, giving the first formula. The second is obtained using the properties of the distance function $r(S) = d^g(S, \mathrm{I}_n)$ (C.4).

Proposition 3.3 (explicit calculation). *The measure of angular diagonality is written*

$$\angle^g(S) = \arccos \frac{\|\text{diag } \ln^g_{\mathrm{I}_n} S\|_{\mathrm{I}_n}}{\|\ln^g_{\mathrm{I}_n} S\|_{\mathrm{I}_n}} = \arccos \|\text{diag grad}^{\,g} r(S)\|_{\mathrm{I}_n}$$

3.1.2 Gradient calculation

In this section, we calculate the gradients to implement our previous framework. This application is differentiable (by composition with the functions arccos, norm, and diag). We calculate directly the differential and deduce the gradients using (C.4). For $S \in S_n^{++}(\mathbb{R})$, $X \in S_n(\mathbb{R})$ and ∇^g the Lévy-Civita connection for g, we get first equality. The definition of Riemannian Hessian gives the second.

$$\mathrm{d}_S\angle^g(X) = -\frac{\langle \nabla^g_X \text{grad}\, r(S), \text{diag grad}^{\,g} r(S)\rangle_{\mathrm{I}_n}}{\sin\angle^g(S)\|\text{diag grad}^{\,g} r(S)\|_{\mathrm{I}_n}} = -\frac{\nabla^g dr\,[\text{diag grad}^{\,g} r(S), X]}{\sin\angle^g(S)\|\text{diag grad}^{\,g} r(S)\|_{\mathrm{I}_n}}$$

Euclidean criterion gradient The definition of gradient gives

$$\text{grad}^{\,F}\angle^F(S) = \frac{\mathrm{I}_n - \text{diag}\, S}{\sin\angle^F(S)\|\mathrm{I}_n - S\|_F\|\mathrm{I}_n - \text{diag}\, S\|_F} - \frac{\cos\angle^F(S)}{\sin\angle^F(S)\|\mathrm{I}_n - S\|_F^2}(\mathrm{I}_n - S)$$

Affine-invariant criterion gradient Calculation (F) gives

$$\text{grad}^{\,F}\angle^{\text{aff}}(S) = \frac{\text{sym}\,[\text{diag } \ln S(\ln S - \text{diag } \ln S)S^{-1}] - S^{-1}\mathrm{d}_S \ln(S\text{diag } \ln SS)\,S^{-1} + \left\langle \frac{\ln S}{\|\ln S\|_{\mathrm{I}_n}}, \mathrm{d}_S \ln(S\text{diag } \ln SS)\right\rangle_{\mathrm{I}_n} \frac{S^{-1}\ln SS^{-1}}{\|\ln S\|_{\mathrm{I}_n}}}{\sin\angle^{\text{aff}}(S)\|\ln S\|_{\mathrm{I}_n}\|\text{diag } \ln S\|_{\mathrm{I}_n}}$$

It suffices to inject this resultant into (5) to obtain the gradient of the simultaneous angular diagonality measure.

Limitation of angle criteria Because of the arccos, this criterion appears to have a huge flaw: we got a $\sin\angle^{\text{aff}}(S)$ in the denominator, which converges towards 0 when S becomes diagonal. Therefore, the gradient algorithm will not converge to the expected solution. In order to avoid this, we have to minimize alternative functions.

3.1.3 Alternative angular ctiteria

(1) Alternative angular criterion $\widetilde{\angle^g}$ We first propose to optimize $\cos\angle^g$ directly. It amounts to directly deleting the arccos. To keep the properties of minimality and cancellation iff $S \in D_n^{++}(\mathbb{R})$, we further transform the criterion with

$$\widetilde{\angle^g} = 1 - \cos\angle^g$$

This alternative criterion has for extrema 0 (when the angle is 0) and 2, but this last case is excluded by the first item in (3.2) because it corresponds to an angle equal to π. By composition, we have $\text{grad}\,\widetilde{\angle^g} = \sin\angle^g \text{grad}\,\angle^g$ deleting the $\sin\angle^g$ in the denominator. This transformation adds critical points, which are all solutions to the problem. So, this optimization problem is equivalent to the initial. Solving the problem still comes down to a simple search for critical points.

(2) Alternative angular criterion $\widehat{\angle^g}$ In the same way, we propose another alternative criterion

$$\widehat{\angle^g} = \sin^2\angle^g$$

We calculate the gradient obtaining $\text{grad}\,\widehat{\angle^g} = 2\sin\angle^g\cos\angle^g\text{grad}\,\angle^g$. Unlike the previous criterion, this one is not equivalent. This transformation adds all the angle points equal to $\pi/2$ as non-solution critical points. If the algorithm converges to x such that $\forall k$, $x\text{Cov}\mathcal{X}_k x^T$ have an angle equal to $\pi/2$, the average performance will be reduced.

(3) Link between both The two alternative transformations are strongly related by the linearization formulas of $\sin^2$ giving

$$\widehat{\angle^g} = \sin^2\angle^g = \frac{1}{2}\left[1 - \cos(2\angle^g)\right] = \frac{1}{2}2\widetilde{\angle^g} \text{ and } 2\frac{\widehat{\angle^g}}{2} = 2\sin^2\frac{\angle^g}{2} = 1 - \cos\angle^g = \widetilde{\angle^g}$$

So their performance is expected to be similar.

It remains to inject this formula in (5) to obtain the gradient of our simultaneous diagonality criteria.

3.2 The distance information: angle/distance criteria

Despite their interest, previous criteria do not distinguish points on the same geodesic joining I_n and S, because they have the same angle at I_n with $D_n(\mathbb{R})$. For example, we consider the points of geodesic perpendicular to $D_n(\mathbb{R})$ at I_n. Those arbitrarily close to I_n will be very close to $D_n(\mathbb{R})$, but their degree of diagonality will remain equal to $\pi/2$. Conversely, those arbitrarily far from $D_n(\mathbb{R})$ will be very far from being diagonal, but their degree of diagonality will also stay equal to $\pi/2$. To compensate for this defect, we propose to ponderate our angle by the Riemannian distance function $r(S) = d^g(S, \mathrm{I}_n)$ and add this information (see figure 2). We add an exponent β to increase or decrease the influence of the distance term. Of course, $\beta = 0$ gives our simple previous angular criterion.

Définition 3.4 (angle/distance criterion)**.** If $\angle^g$ is an angular criterion, we define

$$\forall \beta \in \mathbb{R}_+,\ r^\beta \angle^g : \left\{ \begin{array}{ccc} S_n^{++}(\mathbb{R}) & \longrightarrow & \mathbb{R}_+ \\ S & \longmapsto & r^\beta(S)\angle^g(S) \end{array} \right.$$

.

As the coefficient $r^\beta(S)$ does not depend on Λ_S, we know that the closest diagonal matrix remains the same (that of the angular criterion). Moreover, we find our simple angular criterion with $\beta = 0$. Now, we can calculate the gradient.

3.2.1 Gradient calculation

Using the link between exponential and distance functions (C.4) we get directly

Proposition 3.5. $\mathrm{grad}^{\,g} r^\beta \angle^g(S) = -\beta r^{\beta-2}(S)\angle^g(S) \ln_{\mathrm{I}_n}^g S + r^\beta(S)\mathrm{grad}\,\angle^g(S)$

Like $\beta > 0$ and $r(S) > 0$, the critical points of $r^\beta \angle^g$ are the union of the critical points of $\angle^g$ with points already solutions to the problem. Thus, we have an improved angular criterion containing additional information on the distance from S to I_n. We can apply the gradient algorithm to solve the problem.

Same limitation However, the presence of the term $\mathrm{grad}^{\,g}\angle^g$ leads us to the same problem of divergence of the gradient algorithm. To get rid of the arccos, we propose to replace the angular criteria with their equivalents, giving the criteria $r^\beta \widetilde{\angle^g}$ and $r^\beta \widehat{\angle^g}$.

3.2.2 The $\beta = 2$ case

For $\beta = 2$, our angular criteria give already known divergences (G). For example, the Euclidean criterion $r^2\widehat{\angle^F}$ is the Frobenius divergence introduced in [11] and [12]. Secondly, the affine-invariant criterion $r^2\widehat{\angle^{\mathrm{aff}}}$ is the log-Euclidean divergence introduced in [6] and [22].

Conversely, certain already known divergences are written as alternative angular criteria. For example, the invariant Frobenius divergence of [5] is written as an alternative Euclidean angle/distance criterion. Similarly, the Riemannian divergence is an alternative affine-invariant angle/distance criterion.

3.3 Numerical experiment with the Euclidean metric

We compare our criteria within the previous LU framework (2.1) only on $\mathbb{L}^1 \times \mathbb{U}^1$ because $\mathbb{L}^\times \times \mathbb{U}^1$ gives the same performances (2.4). Within this framework, $r^\beta \widetilde{\angle^F}$ and $r^\beta \widehat{\angle^F}$ gives the tilde-dβangle1-F and hat-dβangle1-F algorithms.

Data simulation We simulate data similarly to previously. In each Monte-Carlo loop, we add noise $\mathcal{B}_k = \boldsymbol{E_k \Delta'_k E_k^T}$ in the same way as $\mathrm{Cov}\mathcal{X}_k$. Finally the noisy covariance matrices are generated $\forall k \in \{1, \cdots, K\}$, $\mathrm{Cov}\mathcal{X}_k = A\mathrm{Cov}\mathcal{S}_k A^T + \sigma^2 \mathcal{B}_k$ where σ is the noise intensity. With whitening, we implement our algorithms and minimize the simultaneous diagonality measure ϕ_{KL} to obtain $\widehat{B}$.

Error calculation We perform the comparisonno with Amari's criterion [35]. Indeed, the Riemannian error does not seem relevant to the LU matrices subset (2.4).

Comparison with riemEucl We use parameters close to our reference [8]: $n = 8$, $K = 20$, and the graphs are drawn according to the intensity of the noise σ. We use gradient algorithm with optimal step and linear search up to Armijo–Goldstein. Convergence issues are avoided by establishing stopping criteria. Having not verified the Lipschitzianity of the gradient, it is essential to specify that if the algorithm does not converge, performances will be poor and lower than that of riemEucl. Stopping criteria are the same for each algorithm: maximum number of iterations set at 10000, minimum gradient norm set at 10^{-6}, minimum step size set at 10^{-10}, and maximum number of steps set at 5000.

Conclusions First, we can see that curves from $\widetilde{\angle^F}$ and $\widehat{\angle^F}$ overlap. As expected, the corresponding algorithm gives the same performance. So, we only present performances from tilde-angle1-F without repeating that it is the same for hat-angle1-F. Comparing performances with different β we get (3):

- for $\beta = 2; 3; 4$, all our angular criteria give lower Amari errors than riemEucl
- $\beta = 3$ gives the best performances
- from $\beta = 6$, angular criteria are less efficient than riemEucl

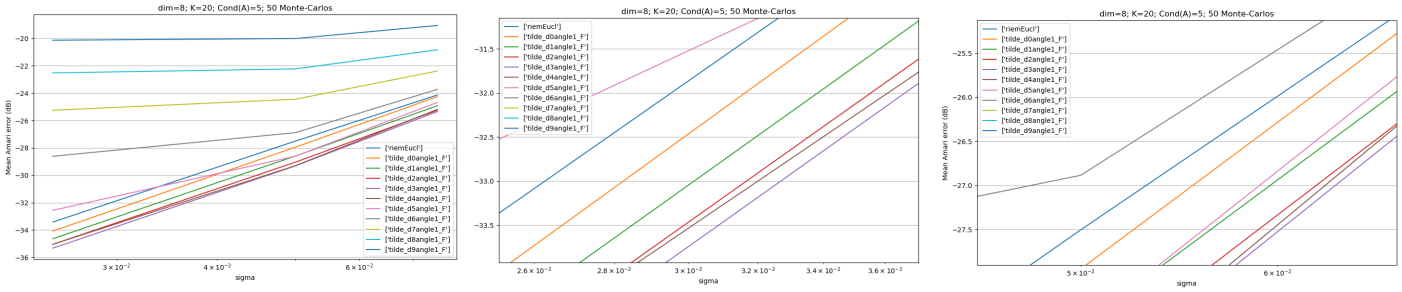

Figure 3: $\sigma \in \{0,25; 0,5; 0,75\}$

With louder noises Taking larger noises, behavior is different:

- $\beta = 2$ performs better than the other angular criteria
- but riemEucl remains the most effective

Our diagonality criteria do not take into account invaiance by permutation and diagonal scaling. In fact, they assign different degrees of diagonality for equivalent solutions. If the gradient algorithm converges to an equivalent solution (from the simulated one), the error will remain large. It is therefore not surprising that increasing noise leads to significant errors. To compensate for this defect, we propose a general method to construct invariant angular diagonality criteria. By assigning the same degree of diagonality to equivalent solutions, the addition of noise should have a lesser impact on the errors.

4 Novel invariant angular criteria

In the previous parts our criteria were not invariant by permutation and diagonal scaling. As algorithms could converges to different equivalent solution, the error could be large even if the solutions were good. Here, we therefore propose a method to construct angular criteria with are invariant by congruence of $\mathcal{P}_n(\mathbb{R})$ and $D_n(\mathbb{R})*$. By giving the same degree of diagonality to equivalent solutions, better performances are expected.

Section 4.1 is devoted to the construction of this criterion. We will the same previous alternative transformations and calculate the corresponding differentials with respect to the differential of ψ. In section 4.2 we add the invariant distance term and calculate the differential. In section 4.3 the numerical experiments show that $\psi(S) = \operatorname{diag} S$ and $\psi(S) = S$ give better than riemEucl and all the previous criterion, including for significant noise.

4.1 Construction of the invariant angular criterion

To construct our invariant criteria, we do not calculate the affine-invariant angle at I_n, but at a point $\psi(S)$ which depends on S. To obtain the simplifications allowing invariance, ψ must be equivariant under the action by congruences of $\mathcal{P}_n(\mathbb{R})$ and $D_n(\mathbb{R})*$. This condition is essential to ensure the invariance.

Définition 4.1 (equivariance)**.** A function $\psi : S_n^{++}(\mathbb{R}) \longrightarrow S_n(\mathbb{R}) \cap \mathrm{GL}_n(\mathbb{R})$ is equivariant under the action by congruence of a subgroup $\mathcal{G} < \mathrm{GL}_n(\mathbb{R})$ if

$$\forall A \in \mathcal{G},\, \forall S \in S_n^{++}(\mathbb{R}),\, \psi(AS\,A^T) = A\psi(S)\,A^T$$

Example $\operatorname{diag} S$, S, $\operatorname{diag}^{-1} S^{-1}$. If $\psi_1, \cdots, \psi_n$ are equivariant functions the alternating products of the type $\psi_1\psi_2^{-1}\psi_3\psi_4^{-1}\psi_5\cdots\psi_n$ are still equivariant.

Now we use ψ to define the following invariant angular criterion.

Définition 4.2 (Invariant angular criterion)**.** We consider a differentiable function $\psi : S_n^{++}(\mathbb{R}) \longrightarrow S_n(\mathbb{R})$ equivariant under the action by congruence of $D_n(\mathbb{R})$ and $\mathcal{P}_n(\mathbb{R})$. We define the invariant angular criterion by

$$i\angle^{\psi}(S) = \arccos \frac{\langle S, \Lambda_S\rangle_{\psi(S)}}{\|S\|_{\psi(S)}\|\Lambda\|_{\psi(S)}}$$

Proposition 4.3. *We have the following properties.*
(1) $i\angle^{\psi}$ is the angle in $\psi(S)$ between the geodesics joining $\psi(S)$ and respectively $\exp^g_{\psi(S)} S$ and $\exp^g_{\psi(S)} \Lambda_S$.
(2) $i\angle^{\psi}$ checks the same properties as the previous angular criteria.
(3) $i\angle^{\psi}$ is invariant under the congruence's action by permutation of $\mathcal{P}_n(\mathbb{R})$ and $D_n(\mathbb{R})^*$.

Proof. (1) by definition of the Riemannian angle. (2) same proof as above. (3) invariance relies on the definition of the affine-invariant metric and on the fact that the action by congruence of $\mathcal{P}_n(\mathbb{R})$ permutes the terms of the diagonal of a symmetric matrix. A direct calculation gives for $S \in S_n^{++}(\mathbb{R})$ and $G \in D_n(\mathbb{R}) \sqcup \mathcal{P}_n(\mathbb{R})$, $i\angle^{\psi}(GSG) = i\angle^{\psi}(S)$. □

Similarly to previously, the Cauchy-Schwarz inequalities gives

Propriétés 4.4 (Explicit calculation)**.** *For $S \in S_n^{++}(\mathbb{R})$ we have*

$$i\angle^{\psi}(S) = \arccos \frac{\langle S, \operatorname{diag} S\rangle_{\psi(S)}}{\|S\|_{\psi(S)}\|\operatorname{diag} S\|_{\psi(S)}}$$

Same limitations Exactly as before we get rid of the arccos with the alternative criteria $\widetilde{i\angle^{\psi}}$ and $\widehat{i\angle^{\psi}}$. By composition $\widetilde{i\angle^{\psi}}$ and $\widehat{i\angle^{\psi}}$ are differentiable. The calculations give As the partial differential with respect to Λ_S is zero, the calculation of the differential in S gives

Proposition 4.5. *For $X \in S_n(\mathbb{R})$ we have*

$$\mathrm{d}_S\widetilde{i\angle^{\psi}}(X) = \frac{\langle \operatorname{diag} S, X\rangle_{\psi(S)} - \langle S\psi(S)^{-1}\operatorname{diag} S + \operatorname{diag} S\psi(S)^{-1}S, \mathrm{d}_S\psi(X)\rangle_{\psi(S)}}{\|S\|_{\psi(S)}\|\operatorname{diag} S\|_{\psi(S)}}$$
$$+ \frac{i\angle^{\psi}(S)\langle \operatorname{diag} S\psi(S)^{-1}\operatorname{diag} S, \mathrm{d}_S\psi(X)\rangle_{\psi(S)}}{\|\operatorname{diag} S\|^2_{\psi(S)}} + \frac{i\angle^{\psi}(S)\left[\langle S\psi(S)^{-1}S, \mathrm{d}_S\psi(X)\rangle_{\psi(S)} - \langle S, X\rangle_{\psi(S)}\right]}{\|S\|^2_{\psi(S)}}$$

and $\mathrm{d}_S\widehat{i\angle^{\psi}}(X) = 2\cos\angle^{\psi}(S)\mathrm{d}_S\widetilde{i\angle^{\psi}}(X)$.

The gradients are deduced using the definition of the gradient for each choice of ψ.

4.2 Distance information

To add the distance information the first idea would be to directly multiply $i\angle^{\psi}$ by the distance between S and $\psi(S)$. But the calculation of the invariant affine distance reveals a ln losing the invarianc. So, instead we propose to multiply by the distance between I_n and $\exp_{\psi(S)} S$. The term in exp makes it possible to remove the ln giving a totally invariant criterion (by the properties of the trace and the equivariance of ψ). The calculation of this term gives the following criterion.

Définition 4.6 (Invariant angle/distance criterion)**.** We consider a differentiable function $\psi : S_n^{++}(\mathbb{R}) \longrightarrow S_n(\mathbb{R})$ equivariant under the action of $\mathcal{P}_n(\mathbb{R})$ and $D_n(\mathbb{R})*$ by congruence. Then for $S \in S_n^{++}(\mathbb{R})$ we define

$$r^{\beta} i\angle^{\psi}(S) = \|S\|_{\psi(S)}^{\beta} i\angle^{\psi}(S) = \|S\|_{\psi(S)}^{\beta} \arccos \frac{\langle S, \Lambda_S \rangle_{\psi(S)}}{\|S\|_{\psi(S)} \|\Lambda\|_{\psi(S)}}$$

Proposition 4.7. *It is an angle/distance criterion invariant under the action by congruence of $\mathcal{P}_n(\mathbb{R})$ and $D_n(\mathbb{R})*$.*

As the distance term does not depend on Λ, the closest diagonal matrix remains the same.

Propriétés 4.8 (Explicit calculation)**.** *For $S \in S_n^{++}(\mathbb{R})$ we have $r^{\beta} i\angle^{\psi}(S) = \|S\|_{\psi(S)}^{\beta} \arccos \frac{\langle S, \operatorname{diag} S \rangle_{\psi(S)}}{\|S\|_{\psi(S)} \|\operatorname{diag} S\|_{\psi(S)}}$.*

This function is differentiable by composition. Calculating the differential gives

$$\mathrm{d}_S r^{\beta} i\angle^{\psi}(X) = \beta \|S\|_{\psi(S)}^{\beta-2} i\angle^{\psi}(S) \left[\langle S, X \rangle_{\psi(S)} - \langle S\psi(S)^{-1} S, \mathrm{d}_S \psi(X) \rangle_{\psi(S)} \right] + \|S\|_{\psi(S)}^{\beta} \mathrm{d}_S i\angle^{\psi}(X)$$

4.3 Numerical experiment

We evaluate the first invariant criteria with $\psi(S) = \operatorname{diag} S$ (iangle-dS). We compare it with riemEucl and the most effective previous angular criteria tilde-d3angle1-F (4). Firstly, for $\sigma < 0.1$, tilde-d3angle1-F maintains the best performance. But for $\sigma > 0,1$ our new invariant angular criterion is more effective than the previous ones, but riemEucl remains slightly better. β decreases performances, event if it is relatively negligible. However, up to $\sigma = 1.1$, $\beta = 0$ gives the best performance, and then it is $\beta = 5$

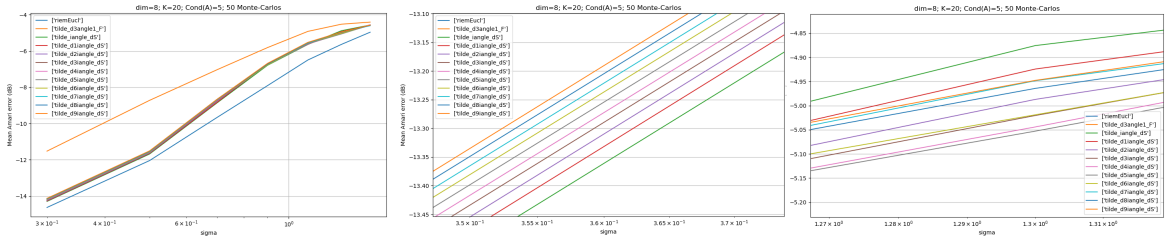

Figure 4: $\sigma \in \{0,3; 0,5; 0,7; 0,9; 1,1; 1,3; 1,5\}$

Now, we do the same with another invariant criterion with $\psi(S) = S$ (denoted iangle-S). Firstly, for $\sigma < 0.1$, tilde-d3angle1-F maintains the best performance. But for $\sigma > 0,1$, performances are better than riemEucl (5). As expected curves for all β overleap, because the distance term is constant.

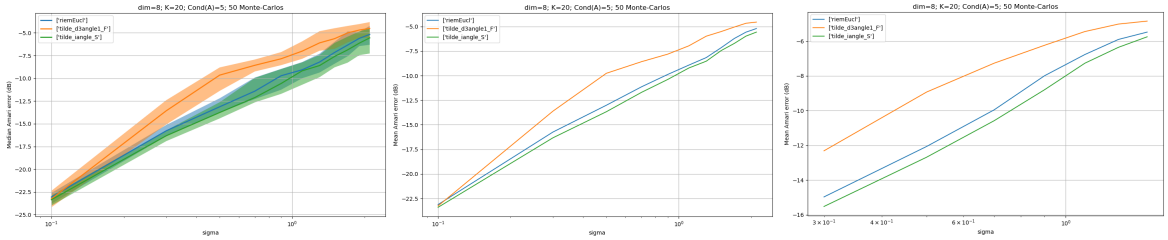

Figure 5: $\sigma \in \{0,1; 0,3; 0,5; 0,7; 0,9; 1,1; 1,3; 1,5; 1,7; 1,9; 2,1\}$

5 Conclusion

5.1 Low noise: the angular criterion

- We defined a new diagonality criteria using Riemannian angle. To do it it is necessary to use a geodesically complete metric such that $D_n^{++}(\mathbb{R})$ is totally geodesic in $S_n^{++}(\mathbb{R})$ (for example the Euclidean and affine-invariant metrics). As the gradient is not defined when S is diagonal, we rather optimize the alternative angular criterion $\widetilde{\angle^g}(S) = 1 - \cos \angle^g(S)$.

  However, angles only gives a partial information. Indeed all the points of the geodesic between I_n and S make the same angle with $D_n^{++}(\mathbb{R})$. In order to add the distance information, we multiplied previous criterion by a distance term, weighted by an exponent β (to be able to control the influence of the new term).

- For $\beta = 2$ we find divergences already used. The Euclidean metric gives the Kullback-Leibler divergence, and the affine-invariant metric gives the log-Euclidean divergence. Conversely, certain already known divergences are written as alternative angular criteria. For example, the invariant Frobenius divergence of is written as an alternative Euclidean angle/distance criterion. Similarly, the Riemannian divergence is an alternative affine-invariant angle/distance criterion.

- For low noise $\sigma < 0.1$ we have developed an optimization framework using LU decomposition. We compare our algorithms with Pham's algorithm (riemEucl). We get better performance for several values of β, the best being $\beta = 3$.

5.2 Higher noise: the invariant diagonality criterion

However, this solution is not completely satisfactory. Indeed our diagonality criteria give different degrees of diagonality for equivalent solutions. This leads to significant errors when the gradient algorithm converges to different equivalent solutions (by permutation and diagonal scaling). The simulated performances are therefore considerably weakened.

- In order to compensate for this, we propose a very general method to construct invariant diagonality criteria (by permutation and diagonal scaling). This method is based on the affine-invariant metric. It consists of measuring the angle in $\psi(S)$ (where ψ is equivariant) between the geodesics joining $\psi(S)$ and respectively $\exp^g_{\psi(S)} S$ and $\exp^g_{\psi(S)} \operatorname{diag} S$. This makes it possible to have the same degree of diagonality for equivalent solutions.

  In the same way as previously, we use previos alternative transformation. As before we add a distance term which is also invariant. We parameterize it again by β.

- For higher noises $\sigma > 0.1$ we obtain better performances. The parameter function $\psi(S) = S$ allows simplifications which give better performance of our riemEucl comparison algorithm.

6 Work to do

Numerical experiments with the invariant affine metric

The first thing to do is to compare performances of angular affine-invariant criteria. For this, we need an approximation of the differential of the matrix's logarithm. This numerical estimate as been treated for example in [3], [4], [13], [15], [20], [26], [27], [34], [33]. A general method is proposed in [27], and the most widespread method is that of [4], extended in [8].

CAT(0) space

With the affine-invariant metric, we know that $S_n^{++}(\mathbb{R})$ is a CAT(0) space (for example, in [38] or [24]). So, thanks to plane trigonometry equality, a method (presented for example in [38]) exists to obtain an upper bound of the Riemannian angle. These inequalities may provide a lot of interesting diagonality criteria. Moreover, the sectional curvature of the manifold is also lower bounded by $-\frac{1}{4}$. Therefore, the equality of hyperbolic trigonometry could give us lower bounds similarly. Combined with the previous increases, we could hope to obtain frameworks giving usable criteria.

Huber's approach

So far, we have chosen angle/distance criteria with constant β. Defining criteria where β would be a function of S might be interesting. The idea would be to adopt a Huber-like approach, varying β according to proximity to I_n. The idea would be to increase or decrease the distance component to improve performance.

Conforming Diagonality Criteria

It could be interesting to consider conforming angular criteria. Indeed, these transformations preserve Riemannian angles, but a conformity term appears when calculating the conformal gradient. The choice of the conformity coefficient would allow us to control the size of the optimization step (by a parameter function) without affecting the calculation of the diagonality criterion. It might be interesting to find out if there are parameter functions that would improve the performance of already existing algorithms.

Second approach

There exists a second approach to the same problem. It uses the same deviation measures to which $\mathrm{GL}_n(\mathbb{R})$ after congruence. For an angular measure $\angle^g$ we no longer seek to minimize the sum of $\angle^g(x\mathrm{Cov}\mathcal{X}_k\, x^T, D_n^{++}(\mathbb{R}))$ but rather the sum of $\angle^g(\mathrm{Cov}\mathcal{X}_k, x^{-1}D_n^{++}(\mathbb{R})x^{-T})$. It would be interesting to use our criteria in this new approach.

Real data

Of course, evaluating these methods on real data will then be necessary.

Open questions from unification research

The methods developed so far have few links between them. Each uses criteria, constraints, and specific optimization methods. Therefore, We cannot estimate the real impact of each choice on the results. For example, [41] and [48] have neither the same criterion, nor the same constraint, nor the same optimization method. Despite the similarity of the problem, the solutions are not comparable. Some work attempts to unify these methods. For example, the α-log-det divergence (from [14]) is a generalisation, giving the S-divergence, the left and the right Kullback-Leibler (with $\alpha = 0; 1$ and -1) [5].

A second type of unification lies in the links between divergences and Riemannian distances. Indeed, if the divergences are ad-hoc defined, distances are intrinsically linked to the geometry of the underlying space. Several works have already explored these questions. For example, many divergences are squared Riemannian distances [5]: Frobenius divergence (Euclidean distance), Frobenius invariant (affine metric invariant on $S_n(\mathbb{R})$), log-Euclidean (affine-invariant metric on $S_n^{++}(\mathbb{R})$), Wasserstein, Riemannian etc. So, it is natural to look for other distance-squared divergences. How characterize them? Do they form a particular space?

Furthermore, [46] goes further by transforming the Bergman divergence into a squared distance. This last article asks many questions. How to know if a divergence can be transformed into a squared distance? How do we find these transformations? What divergences can be transformed into squared distance?

Next, if a divergence is a squared distance, it could be interesting to compare both performances against the approximate joint diagonalization problem. Indeed, the intrinsic nature of the distances would make it possible to deeply exploit the geometry of the space, giving many optimality properties (often lost with the square). For example, a cost distance function (C) allows the gradient algorithm to follow the geodesics, making each iteration geometrically optimal. Moreover, $S_n^{++}(\mathbb{R})$ is a Hadamard space for the invariant affine metric, so its distance function is convex. We can wonder if this convexity could be exploited in a gradient descent.

Open questions from Riemannian geometry

The equations of radial curves [39] directly relate the geometry of a Riemannian submanifold with the gradient and the Hessian of a function. By expressing the Hessian according to the sectional curvature, we ask if we could deduce a notion of iterative convergence speed for gradient or Newton's algorithms. Similarly, the trace of the third radial curves equation gives $\Delta r - \operatorname{tr}\operatorname{Hess}^2 r = -\mathrm{Ric}\,(\operatorname{grad} r, \operatorname{grad} r)$. However, the Laplacian is the trace of the Hessian, and the Ricci curvature gives the variation of a neighborhood of a point transported along a geodesic. We ask if we could deduce another notion of iterative convergence speed. Combined with the results of [21], these equations give artifacts of iterative convergence speed.

A Gradient on LU spaces

Firstly we calculate the Euclidean gradient of ϕ. If the diagonality criterion $\mathcal{D}$ is differentiable, ϕ is a differentiable function and

$$\operatorname{grad}\phi(x) = 2\sum_{k=1}^{K} \operatorname{grad}\mathcal{D}(x\mathrm{Cov}\mathcal{X}_k x^T)x\mathrm{Cov}\mathcal{X}_k \tag{5}$$

Now we calculate the gradient of $\overline{\phi}$. By composition, this function is differentiable

$$\mathrm{d}_{x_L,x_U}\overline{\phi}\,(X_L, X_U) = \mathrm{d}_{x_L x_U}\phi\,(X_L x_U) + \mathrm{d}_{x_L x_U}\phi\,(x_L X_U)$$

The definition of the gradient and the property of trace give

$$\left\langle \operatorname{grad}^{F\times F}\overline{\phi}(x_L, x_U), (X_L, X_U)\right\rangle_{F\times F} = \left\langle \operatorname{grad}\phi(x_L x_U)\, x_U^T, X_L\right\rangle_F + \left\langle x_L^T \operatorname{grad}\phi(x_L x_U), X_U\right\rangle_F$$

By définition of the product metric

$$\left\langle \operatorname{grad}^{F\times F}\overline{\phi}(x_L, x_U), (X_L, X_U)\right\rangle_{F\times F} = \left\langle \begin{bmatrix} \operatorname{grad}\phi(x_L x_U)\, x_U^T \\ x_L^T \operatorname{grad}\phi(x_L x_U) \end{bmatrix}, \begin{bmatrix} X_L \\ X_U \end{bmatrix} \right\rangle_{F\times F}$$

Finaly for $(x_L, x_U) \in \mathbb{L}^1 \times \mathbb{U}^1$ we get the Euclidean gradient of $\overline{\phi}$ on $M_n(\mathbb{R}) \times M_n(\mathbb{R})$

$$\forall (x_L, x_U) \in \mathbb{L}^1 \times \mathbb{U}^1,\ \operatorname{grad}^{F\times F}\overline{\phi}(x_L, x_U) = \begin{bmatrix} \operatorname{grad}\phi(x_L x_U)\, x_U^T \\ x_L^T \operatorname{grad}\phi(x_L x_U) \end{bmatrix} \tag{6}$$

To obtain the gradient on $\mathbb{L} \times \mathbb{U}^0$, we project $\operatorname{grad}^{F\times F}\overline{\phi}$ with the Euclidean orthogonal linear projection. To do this we consider

$$\mathrm{diag} : \begin{cases} M_n(\mathbb{R}) & \longrightarrow & D_n(\mathbb{R}) \\ M & \longmapsto & \operatorname{diag} M \end{cases} \quad \text{and} \quad \mathrm{trig} : \begin{cases} M_n(\mathbb{R}) \times M_n(\mathbb{R}) & \longrightarrow & \mathbb{L}^0 \times \mathbb{U}^0 \\ (M, M') & \longrightarrow & (M_-, M'_+) \end{cases}$$

where $\operatorname{diag} M$ is the diagonal part of M and M_-, M_+ are the strict upper and lower parts of M. As the product between lower (resp. upper) triangular and strict lower (resp. upper) triangular matrices gives a strict lower (resp. upper) triangular matrix, the linear projection $\mathrm{trig} + (\mathrm{diag}\,, 0) : M_n(\mathbb{R}) \longrightarrow \mathbb{L} \times \mathbb{U}^0$ is orthogonal for the Euclidean metric. After projection, equation (6) gives the result.

B Results on the invariant affine metric

We consider the action of $\mathrm{GL}_n(\mathbb{R})$ on $S_n^{++}(\mathbb{R})$ by congruence

$$\pi : \begin{cases} \mathrm{GL}_n(\mathbb{R}) & \longrightarrow & \mathrm{GL}\left[S_n^{++}(\mathbb{R})\right] \\ A & \longmapsto & A \cdot A^T \end{cases}$$

As $\forall A \in \mathrm{GL}_n(\mathbb{R})$, π_A is a linear isomorphism, we have $\forall x \in S_n^{++}(\mathbb{R})$, $\mathbb{T}_x\pi_A = \pi_A$ which justifies that we similarly note the action on $S_n^{++}(\mathbb{R})$ and the action on tangent spaces $S_n(\mathbb{R})$. We equip $S_n^{++}(\mathbb{R})$ with the invariant metric under the action of π

$$\forall x \in S_n^{++}(\mathbb{R}),\ \forall X, Y \in S_n(\mathbb{R}),\ g_x^{\mathrm{aff}}(X, Y) = \mathrm{tr}\left(x^{-1}Xx^{-1}Y\right)$$

Propriétés B.1. *The metric in $S \in S_n^{++}(\mathbb{R})$ is calculated according to the metric in I_n (which is the Euclidean metric) by*

$$g_S^{\mathrm{aff}}(X, Y) = g_{\mathrm{I}_n}^{\mathrm{aff}}\left(\sqrt{S^{-1}}X\sqrt{S^{-1}}, \sqrt{S^{-1}}Y\sqrt{S^{-1}}\right) = g_F^{\mathrm{aff}}\left(\sqrt{S^{-1}}X\sqrt{S^{-1}}, \sqrt{S^{-1}}Y\sqrt{S^{-1}}\right)$$

Proof. By stability of the trace by permutation of the terms and $S^{-1} = \sqrt{S^{-1}}\sqrt{S^{-1}}$ we have

$$g_S^{\mathrm{aff}}(X, Y) = \mathrm{tr}\left(S^{-1}XS^{-1}Y\right) = \mathrm{tr}\left(\sqrt{S^{-1}}X\sqrt{S^{-1}}\sqrt{S^{-1}}Y\sqrt{S^{-1}}\right) = g_{\mathrm{I}_n}^{\mathrm{aff}}(\sqrt{S^{-1}}X\sqrt{S^{-1}}, \sqrt{S^{-1}}Y\sqrt{S^{-1}})$$

□

Proposition B.2. *The geodesic γ of $S_n^{++}(\mathbb{R})$ which verifies $\gamma(0) = S \in S_n^{++}(\mathbb{R})$, $\gamma'(0) = v \in S_n(\mathbb{R})$ is written $\gamma(t) = \sqrt{S}\mathrm{e}^{t\sqrt{S^{-1}}v\sqrt{S^{-1}}}\sqrt{S}$ where* e *is the exponential of the matrices.*

Proof. For $w \in S_n(\mathbb{R})$ the geodesic such that $\gamma(0) = \mathrm{I}_n$ and $\gamma'(0) = w$ is $\forall t \in \mathbb{R}$, $\gamma_{\mathrm{I}_n,w}(t) = \mathrm{e}^{tw}$ where e is the exponential of the matrices (because we are in a Lie group equipped with an affine-invariant metric). As the exponential is a homeomorphism between $S_n(\mathbb{R})$ and $S_n^{++}(\mathbb{R})$, we know that for $S \in S_n^{++}(\mathbb{R})$ its root $\sqrt{S}$ is well defined, and it is still in $S_n^{++}(\mathbb{R})$. By invariance under the action of π we deduce $\sqrt{S} \cdot \gamma_{\mathrm{I}_n,w} = \gamma_{\sqrt{S}\cdot\mathrm{I}_n,\sqrt{S}\cdot w} = \gamma_{\sqrt{S}\sqrt{S}^T,\sqrt{S}w\sqrt{S}}$ which by symmetry of S gives $\sqrt{S} \cdot \gamma_{\mathrm{I}_n,w} = \gamma_{S,\sqrt{S}w\sqrt{S}^T}$. Finally for $v = \sqrt{S}w\sqrt{S}$ we obtain the relation $\gamma_{S,v} = \sqrt{S}\gamma_{\mathrm{I}_n,\sqrt{S}^{-1}v\sqrt{S}^{-1}}\sqrt{S}$ hence the result. □

Proposition B.3. *When defined the associated Riemannian distance d is also affine-invariant.*

Proof. For $S_1, S_2 \in S_n^{++}(\mathbb{R})$ and $A \in \mathrm{GL}_n(\mathbb{R})$ we show that $d(A \cdot S_1, A \cdot S_2) = d(S_1, S_2)$. As we assume that the distance between these points is well defined, we have the existence of the geodesic $\gamma_{S_1,v}$ which joins S_1 to S_2 reparameterized so that $\gamma(1) = S_2$. The definition of the Riemannian distance, the invariance of the geodesics, and the metric successively give

$$\begin{aligned} d^{\mathrm{aff}}(A \cdot S_1, A \cdot S_2) &= \int_0^1 \sqrt{\left\langle \gamma'_{A\cdot S_1,A\cdot v}, \gamma'_{A\cdot S_1,A\cdot v} \right\rangle^{\mathrm{aff}}_{\gamma_{A\cdot S_1,A\cdot v}}} = \int_0^1 \sqrt{\left\langle A\cdot\gamma'_{S_1,v}, A\cdot\gamma'_{S_1,v} \right\rangle^{\mathrm{aff}}_{A\cdot\gamma_{S_1,v}}} \\ &= \int_0^1 \sqrt{\left\langle \gamma'_{S_1,}, \gamma'_{S_1,v} \right\rangle^{\mathrm{aff}}_{\gamma_{S_1,v}}} \end{aligned}$$

i.e. $d^{\mathrm{aff}}(A \cdot S_1, A \cdot S_2) = d^{\mathrm{aff}}(S_1, S_2)$ the associated distance is indeed invariant by the action π . □

Proposition B.4. *$D_n^{++}(\mathbb{R})$ is totally geodesic in $S_n^{++}(\mathbb{R})$ pour Euclidean and affine-invariant metric.*

Proof. If we show that $D_n^{++}(\mathbb{R})$ is totally geodesic in $S_n^{++}(\mathbb{R})$, we will have shown that γ is a geodesic of $D_n^{++}(\mathbb{R})$. Of course, this is the case for the Euclidean metric. It remains to show that $D_n^{++}(\mathbb{R})$ is totally geodesic for the affine-invariant metric. For this, we calculate the second fundamental form. For X, Y vector fields of $S_n^{++}(\mathbb{R})$, we got the Lévy-Civita connection [38]

$$\nabla^{\mathrm{aff}}_X Y = \mathcal{L}_Y X - \frac{YS^{-1}X + XS^{-1}\,, Y}{2} \tag{7}$$

To calculate the second fundamental form, we show that the orthogonal projection on $D_n^{++}(\mathbb{R})$ in $S \in D_n^{++}(\mathbb{R})$ is written

$$\forall X \in \mathbb{T}_S D_n^{++}(\mathbb{R}),\ \mathrm{diag}_S(X) = \sqrt{S}\mathrm{diag}\,(\sqrt{S}^{-1}X\sqrt{S}^{-1})\sqrt{S} \tag{8}$$

Immediately, diag is an Euclidean orthogonal projection on $D_n^{++}(\mathbb{R})$, so it is an orthogonal projection on $D_n^{++}(\mathbb{R})$ in I_n for the affine-invariant metric. By invariance, we deduce

$$\mathrm{diag}_S(Y) = \sqrt{S}\mathrm{diag}_{\mathrm{I}_n}(\sqrt{S}^{-1}Y\sqrt{S}^{-1})\ sqrtS$$

For $Y = \sqrt{S}X\sqrt{S}$ we get the result.

Now, we use it to calculate the second fundamental form in S. Since diagonal matrices commute, for X, $Y \in \mathbb{T}_S D_n(\mathbb{R})$ we have

$$\mathrm{II}_S(X, Y) = \nabla_X Y - \mathrm{diag}_S(\nabla_X Y)$$

The formulas for the connection (7) and the orthogonal projection (8) give

$$\mathrm{II}_S(X, Y) = 0$$

□

C Results on distance functions

Distance functions are essential in Riemannian geometry. They are present in all reference writings such as [38] or [24]. Some writings, such as [39], present Riemannian geometry with particular emphasis on distance functions.

Définition C.1. A distance function of the Riemannian submanifold (M, g) is a function $r : \Omega \subset M \longrightarrow \mathbb{R}$ whose gradient has the norm $\|\mathrm{grad}\, r\| = 1$.

For example, if d is the Riemannian distance of a Riemannian submanifold (M, g) and $x, x_0 \in M$, it is shown in [45] that $x \longmapsto d(x_0, x)$ is a distance function.

Corollaire C.2. $\nabla_{\mathrm{grad}\, r}\mathrm{grad}\, r = 0$

Proof. We consider the function ψ defined by $\forall x \in M$, $\psi(x) = \|\mathrm{grad}\, r\|_x^2$ we have

$$\forall X \in \mathbb{T}_x M,\ \mathrm{d}_x\psi\,(x) = 2\langle \nabla_X \mathrm{grad}\, r, \mathrm{grad}\, \rangle$$

By definition and symmetry of the Hessian, this equality is written

$$\forall X \in \mathbb{T}_x M,\ d_x\psi(x) = 2\mathrm{Hess}\,(X, \mathrm{grad}\, r) = 2\mathrm{Hess}\, r(\mathrm{grad}\, r, X) = 2\langle \nabla_{\mathrm{grad}\, r}\mathrm{grad}\, r, X\rangle$$

But by definition of the gradient we also know that $T\psi(X) = \langle \mathrm{grad}\, \psi, X\rangle$ so by identification we have $2\nabla_{\mathrm{grad}\, r}\mathrm{grad}\, r = \mathrm{grad}\, \psi = \mathrm{grad}\, \|\mathrm{grad}\, r\|^2$. By definition of a distance function, we know that the norm of the gradient of r is constant equal to 1 from where

$$\nabla_{\mathrm{grad}\, r}\mathrm{grad}\, r = \frac{1}{2}\mathrm{grad}\, \|\mathrm{grad}\, r\|^2 = 0$$

□

Conséquence C.3. *the integral curves of the vector field* $\mathrm{grad}\, r$ *are geodesics.*

Proof. By definition of the local flow Φ, for t in the domain of definition of $\Phi_\cdot(x)$ and $x \in M$ we have $\frac{d}{dt}\Phi_t(x) = \mathrm{grad}\, \Phi_t(x)$. We deduce $\nabla_{\Phi'_\cdot}\Phi'_\cdot = \nabla_{\mathrm{grad}\, r}\mathrm{grad}\, r = 0$ (by the previous corollary C.2) which is exactly the definition of a geodesic. The integral curves $\Phi_\cdot(x) : t \longmapsto \Phi_t$ are geodesics. □

Thus, the integral curves of the vector field $-\mathrm{grad}\, r$ are also geodesics. If x_n is a point of one of these geodesic integral curves, its gradient (and its dilated) $-\mathrm{grad}\, r(x_n)$ is a tangent vector to this curve (definition of the local flow); therefore, the exponential gives a new point x_n which is still on the same geodesic integral curve. By induction, the gradient algorithm (with the exponential as retraction) gives a series of points that remain on the geodesic integral curve resulting from the initial point x_0 with initial tangent vector $-\mathrm{grad}\, r(x_0)$.

Corollaire C.4 (link between exponential and distance function)**.** *We consider a Riemannian manifold* (M, g) *and its distance function at* x_0*. For* $x \neq y$ *on the geodesic we have*

$$\exp_x\left[-r(y)\mathrm{grad}^{\,g} r(x)\right] = y$$

which also gives by inversion

$$\ln_x^g y = -r(y)\mathrm{grad}^{\,g} r(y)$$

D Proof of properties (3.2) of angular criterion

Proof. First of all, we can apply it with the projection closed convex theorem with the Euclidean metric, and we can also apply it with the affine-invariant metric because $S_n^{++}(\mathbb{R})$ becomes a CAT(0) space [38].

(1) We know that the definition set of arccos is $[0, \pi]$. As Λ_S is the diagonal matrix closest to S, the angle is necessarily less than $\pi/2$ in absolute value (otherwise, its symmetric with respect to the identity would still be closer to S).

(2) The function is indeed C^2 by composition.

(3) $\Leftarrow$ If S is diagonal we have S and Λ_S on the same geodesic in $D_n(\mathbb{R})$ (by definition of Λ_S) and we deduce that $\angle^g(S) = 0$.

$\Rightarrow$ If the Riemannian angle is zero, it means that

$$\langle \ln_{\mathrm{I}_n}^g S, \ln_{\mathrm{I}_n}^g \Lambda_S\rangle_{\mathrm{I}_n} = \|\ln_{\mathrm{I}_n}^g S\|_{\mathrm{I}_n}\|\ln_{\mathrm{I}_n}^g \Lambda_S\|_{\mathrm{I}_n}$$

By the Cauchy-Schwarz inequality we deduce that $\ln_{\mathrm{I}_n}^g S$ and $\ln_{\mathrm{I}_n}^g \Lambda_S$ are collinear $\exists t \in \mathbb{R}$, $\ln_{\mathrm{I}_n}^g S = t\ln_{\mathrm{I}_n}^g \Lambda_S$. As g is complete the Hopf-Rinow theorem gives the bijectivity of the exponential then $S = \exp_{\mathrm{I}_n}^g(t\ln_{\mathrm{I}_n}^g \Lambda_S)$. As $D_n^{++}(\mathbb{R})$ is totally geodesic for g we have $\Lambda_S \in D_n^{++}(\mathbb{R}) \Longrightarrow \ln_{\mathrm{I}_n}^g \Lambda_S \in D_n(\mathbb{R}) \Longrightarrow S \in D_n^{++}(\mathbb{R})$. □

E Proof explicit formula (3.3) of angular criterion

Proof. As $D_n(\mathbb{R})$ is totally geodesic for g we have $\Lambda \in D_n^{++}(\mathbb{R}) \Longrightarrow \ln_{\mathrm{I}_n}^g \Lambda \in D_n(\mathbb{R})$ so the angle is rewritten

$$\angle^g(S) = \arccos \frac{\langle \mathrm{diag}\ \ln_{\mathrm{I}_n}^g S, \ln_{\mathrm{I}_n}^g \Lambda_S \rangle_{\mathrm{I}_n}}{\| \ln_{\mathrm{I}_n}^g S\|_{\mathrm{I}_n} \| \ln_{\mathrm{I}_n}^g \Lambda_S\|_{\mathrm{I}_n}}$$

Therefore to explicitly calculate $\angle^g$ it is necessary to determine

$$\Lambda_S = \underset{\Lambda \in D_n^{++}(\mathbb{R})}{\mathrm{argmin}}\ \arccos \frac{\langle \mathrm{diag}\ \ln_{\mathrm{I}_n}^g S, \ln_{\mathrm{I}_n}^g \Lambda \rangle_{\mathrm{I}_n}}{\| \ln_{\mathrm{I}_n}^g S\|_{\mathrm{I}_n} \| \ln_{\mathrm{I}_n}^g \Lambda\|_{\mathrm{I}_n}}$$

By decrease of the function arccos and the Cauchy-Schwarz inequality we know that the function

$$\Lambda \longmapsto \arccos \frac{\langle \mathrm{diag}\ \ln_{\mathrm{I}_n}^g S, \ln_{\mathrm{I}_n}^g \Lambda \rangle_{\mathrm{I}_n}}{\| \ln_{\mathrm{I}_n}^g S\|_{\mathrm{I}_n} \| \ln_{\mathrm{I}_n}^g \Lambda\|_{\mathrm{I}_n}}$$

reaches its minimum iff $\ln_{\mathrm{I}_n}^g \Lambda$ and $\mathrm{diag}\ \ln_{\mathrm{I}_n}^g S$ are collinear $\exists t \in \mathbb{R}$, $\ln_{\mathrm{I}_n}^g \Lambda = t\mathrm{diag}\ \ln_{\mathrm{I}_n}^g S$. As $\angle^g(S) \in [0, \pi/2]$ (3.2) we have $t > 0$. By completeness of g we know that $\exp_{\mathrm{I}_n}^g$ is bijective so the last equality is equivalent to

$$\exists t > 0,\ \Lambda = \exp_{\mathrm{I}_n}^g (t\mathrm{diag}\ \ln_{\mathrm{I}_n}^g S)$$

which means that the points which minimize this function are those of the geodesic which joins I_n and $\mathrm{diag}\ \ln_{\mathrm{I}_n}^g S$. When we inject this equality into the formula of $\angle^g$, the $\ln_{\mathrm{I}_n}^g$ are simplified with the $\exp_{\mathrm{I}_n}^g$, then the $t > 0$ and the norms are simplified by bilinearity of the metric and homogeneity of the norm. We get the result. □

F Gradient affine-invariant angular criterion calculation

To use the compatibility of the metric with the Lévy-Civita connection we rewrite

$$\angle^{\mathrm{aff}}(S) = \sqrt{\langle S\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S) S, \mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S) \rangle_S}$$

By composition, by compatibility of the metric and by definition of the affine-invariant metric we have for $S \in S_n^{++}(\mathbb{R})$ and $X \in S_n(\mathbb{R})$

$$\mathrm{d}_S \angle^{\mathrm{aff}}(X) = -\frac{\left\langle \mathrm{sym}\left[S\mathrm{diag}^{\,2}\mathrm{grad}^{\mathrm{aff}} r(S)\right], X\right\rangle_S + \left\langle \nabla_X^{\mathrm{aff}} \mathrm{grad}^{\mathrm{aff}} r(S), S\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S) S\right\rangle_S}{\sin \angle^{\mathrm{aff}}(S) \|\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S)\|_{\mathrm{I}_n}}$$

By Hessian symmetry we have

$$\mathrm{d}_S \angle^{\mathrm{aff}}(X) = -\frac{\left\langle \mathrm{sym}\left[S\mathrm{diag}^{\,2}\mathrm{grad}^{\mathrm{aff}} r(S)\right], X\right\rangle_S + \left\langle \nabla^{\mathrm{aff}}_{S\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S) S} \mathrm{grad}^{\mathrm{aff}} r(S), X\right\rangle_S}{\sin \angle^{\mathrm{aff}}(S) \|\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S)\|_{\mathrm{I}_n}}$$

By definition of the gradient

$$\mathrm{grad}^{\mathrm{aff}} \angle^{\mathrm{aff}}(S) = -\frac{\mathrm{sym}\left[S\mathrm{diag}^{\,2}\mathrm{grad}^{\mathrm{aff}} r(S)\right] + \nabla^{\mathrm{aff}}_{S\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S) S} \mathrm{grad}^{\mathrm{aff}} r(S)}{\sin \angle^{\mathrm{aff}}(S) \|\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S)\|_{\mathrm{I}_n}}$$

The link between the distance function and the exponential (C.4) gives

$$\mathrm{grad}^{\mathrm{aff}} \angle^{\mathrm{aff}}(S) = -\frac{\mathrm{sym}\left(S\mathrm{diag}^{\,2} \ln S\right) + \| \ln S\|_{\mathrm{I}_n} \nabla^{\mathrm{aff}}_{S\mathrm{diag}\ \ln SS} \frac{\ln S}{\| \ln S\|_{\mathrm{I}_n}}}{\sin \angle^{\mathrm{aff}}(S) \| \ln S\|_{\mathrm{I}_n} \|\mathrm{diag}\,\mathrm{grad}^{\mathrm{aff}} r(S)\|_{\mathrm{I}_n}}$$

The calculation of the Lévy-Civita connection for the affine-invariant metric, then the calculation of the Lie derivative gives

$$\begin{aligned}
\nabla^{\mathrm{aff}}_{S\mathrm{diag}\ \ln SS} \frac{\ln S}{\| \ln S\|_{\mathrm{I}_n}} &= \mathcal{L}_{S\mathrm{diag}\ \ln SS} \frac{\ln S}{\| \ln S\|_{\mathrm{I}_n}} - \frac{\mathrm{sym}\left(\ln S\mathrm{diag}\ \ln SS\right)}{\| \ln S\|_{\mathrm{I}_n}} \\
&= \frac{\mathrm{d}_S \ln\left(S\mathrm{diag}\ \ln SS\right) - \left\langle \frac{\ln S}{\| \ln S\|_{\mathrm{I}_n}}, \mathrm{d}_S \ln\left(S\mathrm{diag}\ \ln SS\right)\right\rangle \ln S - \mathrm{sym}\left(\ln S\mathrm{diag}\ \ln SS\right)}{\| \ln S\|_{\mathrm{I}_n}}
\end{aligned}$$

Finally we obtain

$$\mathrm{grad}^{\mathrm{aff}} \angle^{\mathrm{aff}}(S) = \frac{\mathrm{sym}\left[S\mathrm{diag}\ \ln S(\ln S - \mathrm{diag}\ \ln S)\right] - \mathrm{d}_S \ln\left(S\mathrm{diag}\ \ln SS\right) + \left\langle \frac{\ln S}{\| \ln S\|_{\mathrm{I}_n}}, \mathrm{d}_S \ln\left(S\mathrm{diag}\ \ln SS\right)\right\rangle_{\mathrm{I}_n} \frac{\ln S}{\| \ln S\|_{\mathrm{I}_n}}}{\sin \angle^{\mathrm{aff}}(S) \| \ln S\|_{\mathrm{I}_n} \|\mathrm{diag}\ \ln S\|_{\mathrm{I}_n}}$$

G Proof of $\beta = 2$ case.

$r^2\widehat{\angle^F}$ is the Frobenius divergence

Propriétés G.1 (Frobenius). *$r^2\widehat{\angle^F}$ is the Frobenius divergence.*

Proof. For $S \in S_n^{++}(\mathbb{R})$ the bilinearity of the metric, then the properties of diag allow to rewrite the Frobenius divergence

$$\begin{aligned}\mathcal{D}_F(S) &= \|S - \operatorname{diag} S\|_F^2 = \|(S - \mathrm{I}_n) - \operatorname{diag}(S - \mathrm{I}_n)\|_F^2 \\ &= \|S - \mathrm{I}_n\|_F^2 - 2\langle S - \mathrm{I}_n, \operatorname{diag}(S - \mathrm{I}_n)\rangle_F + \|\operatorname{diag} S - \mathrm{I}_n\|_F^2 \\ &= \|S - \mathrm{I}_n\|_F^2 - 2\|\operatorname{diag} S - \mathrm{I}_n\|_F^2 + \|\operatorname{diag} S - \mathrm{I}_n\|_F^2 = r^2\widehat{\angle^F}(S)\end{aligned}$$

□

$r^2\widehat{\angle^{\mathrm{aff}}}$ is the log-Euclidean divergence

Propriétés G.2 (log-Euclidean). *$r^2\widehat{\angle^{\mathrm{aff}}}$ is the log-Euclidean divergence.*

Proof. For $S \in S_n^{++}(\mathbb{R})$ the bilinearity of the metric, then the properties of diag allow to rewrite the log-Euclidean divergence

$$\begin{aligned}\mathcal{D}_{LE}(S) &= \|\ln S - \operatorname{diag}\ln S\|_{\mathrm{I}_n}^2 = \|\ln S\|_{\mathrm{I}_n}^2 - 2\langle \ln S, \operatorname{diag}\ln S\rangle_{\mathrm{I}_n} + \|\operatorname{diag}\ln S\|_{\mathrm{I}_n}^2 \\ &= \|\ln S\|_{\mathrm{I}_n}^2 - 2\|\operatorname{diag}\ln S\|_{\mathrm{I}_n}^2 + \|\operatorname{diag}\ln S\|_{\mathrm{I}_n}^2 = r^2\widehat{\angle^{\mathrm{aff}}}(S)\end{aligned}$$

□

Invariant Frobenius divergence as alternative angular criterion

Propriétés G.3 (invariant Frobenius divergence). *The invariant Frobenius divergence is an alternative Euclidean angle/distance criterion*

$$\begin{aligned}\mathcal{D}_{FI}(S) &= r_F^2(S\operatorname{diag}^{-1}S)\cos\angle_{\mathrm{I}_n}^F\left[\ln_{\mathrm{I}_n}^F(\operatorname{diag}^{-1}SS), \ln_{\mathrm{I}_n}^F(S\operatorname{diag}^{-1}S)\right] \\ &= r_F^2(\operatorname{diag}^{-1}SS)\cos\angle_{\mathrm{I}_n}^F\left[\ln_{\mathrm{I}_n}^F(\operatorname{diag}^{-1}SS), \ln_{\mathrm{I}_n}^F(S\operatorname{diag}^{-1}S)\right]\end{aligned}$$

Proof. By definition of the invariant Frobenius divergence and the properties of the trace

$$\mathcal{D}_{FI}(S) = \langle \operatorname{diag}^{-1}SS - \mathrm{I}_n, S\operatorname{diag}^{-1}S - \mathrm{I}_n\rangle_F$$

By definition of $\ln^F$ we have

$$\mathcal{D}_{FI}(S) = \langle \ln_{\mathrm{I}_n}^F(\operatorname{diag}^{-1}SS), \ln_{\mathrm{I}_n}^F(S\operatorname{diag}^{-1}S)\rangle_F$$

By definition of the Euclidean angular criterion

$$\mathcal{D}_{FI}(S) = \|\ln_{\mathrm{I}_n}^F(\operatorname{diag}^{-1}SS)\|_F\|\ln_{\mathrm{I}_n}^F(S\operatorname{diag}^{-1}S)\|_F\angle^F\left[\ln_{\mathrm{I}_n}^F(\operatorname{diag}^{-1}SS), \ln_{\mathrm{I}_n}^F(S\operatorname{diag}^{-1}S)\right]$$

From the properties of the trace we know that $\|\ln_{\mathrm{I}_n}^F(\operatorname{diag}^{-1}SS)\|_F = \|\ln_{\mathrm{I}_n}^F(S\operatorname{diag}^{-1}S)\|_F = r^F(S\operatorname{diag}^{-1}S) = r^F(\operatorname{diag}^{-1}SS)$ □

Riemannian divergence as alternative angular criterion

Propriétés G.4 (Riemannian divergence). *The Riemannian divergence is an alternative affine-invariant angle/distance criterion*

$$\mathcal{D}_R(S) = r_{\mathrm{aff}}^2(S\Lambda_S^{R^{-1}})\cos\angle_{\mathrm{I}_n}^{\mathrm{aff}}\left[\ln(S\Lambda_S^{R^{-1}}), \ln(\Lambda_S^{R^{-1}}S)\right] = r_{\mathrm{aff}}^2(\Lambda_S^{R^{-1}}S)\cos\angle_{\mathrm{I}_n}^{\mathrm{aff}}\left[\ln(S\Lambda_S^{R^{-1}}), \ln(\Lambda_S^{R^{-1}}S)\right]$$

Proof. For $S \in S_n^{++}(\mathbb{R})$ and Λ_S^R its closest (positive definite) diagonal matrix (for the affine-invariant distance), the properties of the matrix ln and the trace give

$$\mathcal{D}_R(S) = d^2_{\text{aff}}(S, \Lambda_S^R) = \left\| \ln\left(\sqrt{\Lambda_S^R}^{-1} S \sqrt{\Lambda_S^R}^{-1} \right) \right\|_F^2 = \left\langle \ln(S\Lambda_S^{R^{-1}}), \ln(\Lambda_S^{R^{-1}} S) \right\rangle_F$$

Since the matrix ln is also $\ln_{\mathrm{I}_n}^{\text{aff}}$ and since the affine-invariant metric in I_n is equal to the Euclidean metric, we have

$$\mathcal{D}_R(S) = \left\| \ln(S\Lambda_S^{R^{-1}}) \right\|_F \left\| \ln(\Lambda_S^{R^{-1}} S) \right\|_F \cos \angle_{\mathrm{I}_n}^{\text{aff}} \left[\ln(S\Lambda_S^{R^{-1}}), \ln(\Lambda_S^{R^{-1}} S) \right]$$

where $\angle_{\mathrm{I}_n}^{\text{aff}} \left[\ln(S\Lambda_S^{R^{-1}}), \ln(\Lambda_S^{R^{-1}} S) \right]$ is again the affine-invariant angle at I_n between the minimizing geodesic which joins I_n to $\ln(S\Lambda_S^{R^{-1}})$ and the minimizing geodesic which joins I_n to $\ln(\Lambda_S^{R^{-1}} S)$. This equality can be rewritten with the affine-invariant distance

$$\mathcal{D}_R(S) = d^{\text{aff}}(\mathrm{I}_n, S\Lambda_S^{R^{-1}}) d^{\text{aff}}(\mathrm{I}_n, \Lambda_S^{R^{-1}} S) \cos \angle_{\mathrm{I}_n}^{\text{aff}} \left[\ln(S\Lambda_S^{R^{-1}}), \ln(\Lambda_S^{R^{-1}} S) \right]$$

As the affine-invariant metric is also invariant under the action of $\mathrm{GL}_n(\mathbb{R})$ by translation to the left and the right, we know that it is the same for the distance (B.3). We can deduce $d^{\text{aff}}\left(\mathrm{I}_n, S\Lambda_S^{R^{-1}}\right) = d^{\text{aff}}(\Lambda_S^R, S)$ and $d^{\text{aff}}(\mathrm{I}_n, \Lambda_S^{R^{-1}} S) = d^{\text{aff}}\left(\Lambda_S^R, S\right)$ hence the result. □

References

[1] P. Ablin, J. Cardoso, and A. Gramfort. Beyond pham's algorithm for joint diagonalization. *Proc ESANN*, 2019.

[2] P.-A. Absil and K. A. Gallivan. Joint diagonalization on the oblique manifold for independent component analysis. In S. Acoustics and . I. . P. Signal Processing, editors, *IEEE International Conference on. T. 5*, 2006.

[3] A. H. Al-Mohy. A more accurate briggs method for the logarithm. *Numerical Algorithms*, 59.3:393–402, 2012.

[4] A. H. Al-Mohy, N. J. Higham, and S. D. Relton. Computing the fréchet derivative of the matrix logarithm and estimating the condition number. *SIAM Journal on Scientific Computing*, 35.4:394–410, 2013.

[5] K. Alyani, M. Congedo, and M. Moakher. Diagonality measures of hermitian positive-definite matrices with application to the approximate joint diagonalization problem. *Linear Algebra and its Applications*, 19(3):290–320, 2016.

[6] S.-I. Amari. Natural gradient works efficiently in learning. 1998.

[7] A. Belouchrani and al. A blind source separation technique using second-order statistics. 1997.

[8] F. Bouchard. *Géométrie and optimisation riemannienne pour la diagonalisation conjointe : application à la séparation de sources d'électroencéphalogrammes*. PhD thesis, ISCE, 11 2018. M. Congedo, J. Malick (supervisors).

[9] F. Bouchard and al. Approximate joint diagonalization with riemannian optimization on the general linear group. 2016.

[10] F. Bouchard, J. Malick, and M. Congedo. Riemannian optimization and approximate joint diagonalization for blind source separation. 2018.

[11] J.-F. Cardoso and A. Souloumiac. Blind beamforming for non gaussian signals. *IEEE Proceedings-F*, (140.6):362–370, 1993.

[12] J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM journal on matrix analysis and applications*, (17.1), 1996.

[13] J. R. Cardoso and F. S. Leite. Theoretical and numerical considerations about padé approximants for the matrix logarithm. *Linear Algebra and its Applications*, (330.1):31–42, 2001.

[14] Z. Chebbi and M. Moakher. Means of hermitian positive-definite matrices based on the log-determinant-divergence function. *Linear Algebra and its Applications*.

[15] S. H. Cheng and al. Approximating the logarithm of a matrix to specified accuracy. *SIAM Journal on Matrix Analysis and Applications*, 22.4:1112–1125, 2001.

[16] P. Comon. Independent component analysis a new concept? *Signal processing*, (36.3):287–314, 1994.

[17] P. Comon and C. Jutten. *Handbook of Blind Source Separation : Independent Component Analysis and Applications*. 2010.

[18] M. Congedo, C. Gouy-Pailler, and C. Jutten. On the blind source separation of human electroencephalogram by approximate joint diagonalization of second order statistics. *Clinical Neurophysiology*, 119(12):2677–2686, 2008.

[19] M. Congedo and D.-T. Pham. Least-squares joint diagonalization of a matrix set by a congruence transformation. 2009.

[20] L. Dieci, B. Morini, and A. Papini. Computational techniques for real logarithms of matrices. *SIAM Journal on Matrix Analysis and Applications*, (17.3):570–593, 1996.

[21] O. P. Ferreira, M. S. Louzeiro, and L. F. Prudente. Gradient method for optimization on riemannian manifolds with lower bounded curvature. *arXiv:1806.02694v1 [math.OC]*, 2018.

[22] P. Fillard and al. *A Riemannian framework for the processing of tensor-valued images*. 2005.

[23] N. Flury and W. Gautschi. An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. 1986.

[24] S. Gallot, D. Hulin., and J. Lafontaine. *Riemannian geometry*. Springer, 3 edition, 2004.

[25] R. Herbin. Cours d'analyse numérique de licence, université d'aix-marseille, 2010.

[26] N. J. Higham. Evaluating padé approximants of the matrix logarithm. *SIAM Journal on Matrix Analysis and Applications*, 22.4:1126–1135, 2001.

[27] N. J. Higham. Functions of matrices : theory and computation. *SIAM*, 2008.

[28] W. Huang, P.-A. Absil, and K. Gallivan. A riemannian bfgs method for nonconvex optimization problems. 2016.

[29] M. Joho. Newton method for joint approximate diagonalization of positive definite hermitian matrices. 2008.

[30] M. Joho and H. Mathis. Joint diagonalization of correlation matrices by using gradient methods with application to blind signal separation. 2002.

[31] M. Joho and K. Rahbar. Joint diagonalization of correlation matrices by using newton methods with application to blind signal separation. 2002.

[32] C. Jutten and J. Herault. Blind separation of sources, part i : An adaptive algorithm based on neuromimetic architecture. *Signal processing*, (24.1):1–10, 1991.

[33] C. S. Kenney and A. J. Laub. Condition estimates for matrix functions. *SIAM journal on matrix analysis and applications*, 10.2:191–209, 1989.

[34] C. S. Kenney and A. J. Laub. A schur-fréchet algorithm for computing the logarithm and exponential of a matrix. *SIAM journal on matrix analysis and applications*, 19.3:640–663, 1998.

[35] E. Moreau and O. Macchi. A one stage self-adaptive algorithm for source separation. volume 3. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2019-119.pdf, https://hal.archives-ouvertes.fr/hal-01936887v1, https://arxiv.org/abs/1811.11433, 1994.

[36] E. Ollila, D. E. Tyler, V. Koinunen, and H. V. Poor. Complex elliptically symmetric distributions: Survey, new results and applications. *IEEE Transactions on Signal Processing*, 2012.

[37] R. M. P.-A. Absil and R. Sepulchre. Optimization algorithms on matrix manifolds. 2008.

[38] F. Paulin. Groupes and géométries, 2014.

[39] P. Pettersen. *Riemannian geometry*. Springer, 3 edition, 2016.

[40] D. T. Pham. Blind separation of instantaneous mixture of sources via an independent component analysis. 1996.

[41] D.-T. Pham. Joint approximate diagonalization of positive definite hermitian matrices. 2000.

[42] D. T. Pham and J.-F. Cardoso. Blind separation of instantaneous mixtures of nonstationary sources. *IEEE Transactions on Signal Processing*, (49.9):1837–1848, 2001.

[43] D.-T. Pham and M. Congedo. Least square joint diagonalization of matrices under an intrinsic scale constraint. 2009.

[44] K. Rahbar and J. P. Reilly. Geometric optimization methods for blind source separation of signals. 2000.

[45] Sakai. *Riemannian geometry*. American Mathematical Society, 3 edition, 1992.

[46] S. Sra. Positive definite matrices and the s-divergence. *arXiv:1110.1773v4 [math.FA]*, 2013.

[47] F. J. Theis, T. P. Cason, and P.-A. Absil. Soft dimension reduction for ica by joint diagonalization on the stiefel manifold. 2009.

[48] P. Tichavsk'y and A. Yeredor. Fast approximate joint diagonalization incorporating weight matrices. 2009.

[49] K. Todros and J. Tabrikian. Fast approximate joint diagonalization of positive definite hermitian matrices. In I. I. C. on. T. 3. ICASSP, editor, *Acoustics, Speech and Signal Processing*, 2007.

[50] D. E. Tyler. A distribution-free m-estimator of multivariate scatter. 1987.