JMLR: Workshop and Conference Proceedings 63:430–444, 2016 ACML 2016

Geometry-aware Stationary Subspace Analysis

Inbal Horev INBAL@MS.K.U-TOKYO.AC.JP
Department of Complexity Science and Engineering, University of Tokyo
5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan

Florian Yger FLORIAN.YGER@DAUPHINE.FR
LAMSADE Université Paris-Dauphine
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France

Masashi Sugiyama SUGI@K.U-TOKYO.AC.JP
Department of Complexity Science and Engineering, University of Tokyo
5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

In many real-world applications data exhibits *non-stationarity*, i.e., its distribution changes over time. One approach to handling non-stationarity is to remove or minimize it before attempting to analyze the data. In the context of brain computer interface (BCI) data analysis this is sometimes achieved using *stationary subspace analysis* (SSA). The classic SSA method finds a matrix that projects the data onto a stationary subspace by optimizing a cost function based on a matrix divergence. In this work we present an alternative method for SSA based on a symmetrized version of this matrix divergence. We show that this frames the problem in terms of distances between symmetric positive definite (SPD) matrices, suggesting a geometric interpretation of the problem. Stemming from this geometric viewpoint, we introduce and analyze a method which utilizes the geometry of the SPD matrix manifold and the invariance properties of its metrics. Most notably we show that these invariances alleviate the need to whiten the input matrices, a common step in many SSA methods which often introduces error. We demonstrate the usefulness of our technique in experiments on both synthetic and real-world data.

Keywords: Stationary subspace analysis, dimensionality reduction, Riemannian geometry, SPD manifold, Grassmann manifold

1. Introduction

A common assumption in statistical modeling is that the distribution of observed data does not change over time, i.e., that it is *stationary*. In most cases it is this assumption of stationarity which allows results to be effectively generalized from the sample to the population. When the stationarity assumption is violated, as is often the case in real-world applications such as speech enhancement (Cohen and Berdugo, 2001) or neurological data analysis (Samek et al., 2012), specialized machine learning methods must be developed in order to maintain adequate prediction capabilities.

A relatively well-studied non-stationary setting is covariate-shift (Shimodaira, 2000), in which the input distribution changes but the conditional distribution of the outputs does not. The problem of covariate-shift has received growing attention in recent years, and

many theoretical and practical aspects have been addressed (see Sugiyama and Kawanabe (2012) for an in-depth exploration of this topic).

These works typically do not aim to remove or reduce non-stationarity in the data, but rather they try to cope with its existence. A different approach is to remove, or minimize, any existing non-stationarities before attempting to analyze the collected data. In the context of brain computer interface (BCI) data analysis, two such note-worthy methods are stationary subspace analysis (SSA) (von Bünau et al., 2009a) and stationary common spacial patterns (sCSP) (Wojcikiewicz et al., 2011).

Similar in spirit to independent component analysis (ICA) (Hyvärinen et al., 2004), SSA statistically models the data as a mixture of stationary and non-stationary signals. Unlike ICA, however, the signals are not assumed to be independent of each other. The data is first split into (possibly overlapping) time frames called epochs. Then a projection matrix is found by optimizing a cost function based on the divergence between distributions in various epochs.

Although SSA is essentially an unsupervised method, variations of it exist which are useful for supervised tasks such as classification (Samek et al., 2012). These methods attempt to remove non-stationarity while keeping the discriminative inter-class variations intact. A different supervised method, by now quite a standard step for classification tasks in BCI systems, is sCSP. Its goal is to project the data onto a subspace in which the various data classes are more separable. The sCSP method directs this subspace towards a stationary subspace by means of regularization.

In this work we present another approach to stationary subspace analysis, focusing for the moment on the unsupervised setting. Unlike SSA, the inputs to our algorithm are not the raw signals themselves, but rather covariance matrices, computed from the signals (or from features based on the signals) using one of the many existing covariance estimators (e.g., Ledoit and Wolf (2004)).

Covariance matrices have gained increasing attention in recent years, and are now commonly used in many machine learning and signal processing applications such as computer vision applications (Tuzel et al., 2006), brain imaging (Pennec et al., 2006) and BCI data analysis (Barachant et al., 2013). Their rich mathematical structure has been extensively studied (Bhatia, 2009), and advances in optimization methods on matrix manifolds in recent years have motivated the development of geometric methods for various tasks such as dictionary learning (Cherian and Sra, 2014), metric learning (Kusner et al., 2014) and dimensionality reduction (Fletcher et al., 2004).

Many methods, and among them SSA, whiten the input covariance matrices as a way to handle correlation between input signals (Hyvärinen et al., 2004). However, as the signal covariances are often erroneously estimated (due to noise or small sample size, for example), this introduces spurious errors. In this work we propose a method that, using the geometric properties of the symmetric positive definite (SPD) matrix manifold and its metrics, not only to produce better stationary subspace estimation, but also to alleviate the need to whiten the input covariance matrices.

2. Geometry-aware stationary subspace analysis

As discussed in the introduction, the task of extracting the stationary part from an observed mixture of stationary and non-stationary signals is essential in various applications. In this section we present our approach to this problem. We name it *geometry-aware SSA* (gaSSA) since we utilize the geometric properties of covariance matrices.

To find a stationary subspace, SSA uses a cost function which is based on a matrix divergence. As a first step we suggest a formulation that uses a symmetrized version of the same matrix divergence. We then show that this symmetrized matrix divergence is a distance between SPD matrices and offer a geometric interpretation to the problem of SSA. We end this section with a theoretical analysis which leads to an elegant, simplified form for gaSSA.

2.1. Stationary subspace analysis

We begin with a formal statement of the problem. To this end we provide a review of the original SSA model and framework (von Bünau et al., 2009a): Let $x(t) \in \mathbb{R}^D$ be a vector of D input signals, composed of m stationary sources $s^{\mathfrak{s}}(t) = [s_1(t), \ldots, s_m(t)]^\top$ ($\mathfrak{s}$-sources) and $D-m$ non-stationary sources $s^{\mathfrak{n}}(t) = [s_{m+1}(t), \ldots, s_D(t)]^\top$ ($\mathfrak{n}$-sources), mixed by a linear mixing transformation,

$$x(t) = As(t) = \left[A^{\mathfrak{s}} A^{\mathfrak{n}}\right] \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}, \tag{1}$$

where $A \in GL_D(\mathbb{R})$, the general linear group of size D over $\mathbb{R}$, i.e., the set of all $D \times D$ invertible matrices with entries in $\mathbb{R}$. The spaces spanned by the column vectors $A^{\mathfrak{s}}$ and $A^{\mathfrak{n}}$ are referred to as the $\mathfrak{s}$-space and $\mathfrak{n}$-space, respectively.

The SSA model makes relatively few assumptions on the $\mathfrak{s}$- and $\mathfrak{n}$-sources. First, the $\mathfrak{s}$-sources are stationary only in the weak (or wide) sense (Kantz and Schreiber, 2004). That is, their first and second moments are required to be constant in time. For the $\mathfrak{n}$-sources, their first two moments may vary between epochs of length T denoted by $\tau_i = [t_0(i), \ldots, t_0(i)+T]$, where $t_0(i)$ is the start time of the i-th epoch. The sources do not necessarily follow a Gaussian distribution, but non-stationarities are assumed to be visible in the first two moments. Furthermore, this model does not assume that the sources are independent, namely, their covariance matrix is given by

$$\Sigma(\tau_i) = \mathbb{E}\left[s(\tau_i)s(\tau_i)^\top\right] = \begin{pmatrix} \Sigma^{\mathfrak{s}} & \Sigma^{\mathfrak{sn}}(\tau_i) \\ \Sigma^{\mathfrak{sn}}(\tau_i)^\top & \Sigma^{\mathfrak{n}}(\tau_i) \end{pmatrix}, \tag{2}$$

where $\Sigma^{\mathfrak{s}} \in \mathbb{R}^{m\times m}$, $\Sigma^{\mathfrak{n}} \in \mathbb{R}^{(D-m)\times(D-m)}$ and $\Sigma^{\mathfrak{sn}} \in \mathbb{R}^{m\times(D-m)}$. Since $\Sigma^{\mathfrak{s}}$ is time independent we have dropped the notation (τ_i).

The goal of SSA is to find a de-mixing transformation $\widehat{A}^{-1}$ that separates the $\mathfrak{s}$-sources from the $\mathfrak{n}$-sources. This matrix $\widehat{A}^{-1}$ is not unique, but rather undetermined up to scaling, sign and linear transformations within each of the $\mathfrak{s}$- and $\mathfrak{n}$-spaces. So, w.l.o.g., the data may be centered and whitened such that the $\mathfrak{s}$-sources have a zero mean and a diagonal covariance matrix with unit variance.[1] Put differently, the de-mixing matrix is written as

1. This is also common practice in ICA (Hyvärinen et al., 2004)

$\widehat{A}^{-1} = \widehat{B}Z$ where $Z = \mathrm{Cov}(x)^{-1/2}$ is a whitening matrix created by a covariance estimator $\mathrm{Cov}(\cdot)$ (in this case it is the empirical estimator) and $\widehat{B} \in \mathcal{O}_D = \left\{V \in \mathbb{R}^{D\times D} \ : \ V^\top V = \mathbb{I}\right\}$, the set of all $D \times D$ orthogonal matrices.

To find the matrix $\widehat{B}$, the signals are split into N epochs $\tau_1, \ldots, \tau_N$ of length T. Each epoch is characterized by its empirical mean $\widehat{\mu}_i$ and covariance $\widehat{\Sigma}_i$. Then for each epoch, the mean and covariance of the $\mathfrak{s}$-sources may be written as

$$\widehat{\mu}_i^{\mathfrak{s}} = \mathbb{I}_D^m \widehat{B} Z \widehat{\mu}_i, \qquad \widehat{\Sigma}_i^{\mathfrak{s}} = \mathbb{I}_D^m \widehat{B} Z \widehat{\Sigma}_i \left(\mathbb{I}_D^m \widehat{B} Z\right)^\top, \tag{3}$$

where $\mathbb{I}_D^m$ is the $D \times D$ identity matrix, truncated to the first m columns. Since the true $\mu^{\mathfrak{s}}$ and $\Sigma^{\mathfrak{s}}$ are by definition stationary, the matrix $\widehat{B}$ is the one which achieves minimal variation of $\widehat{\mu}_i^{\mathfrak{s}}$ and $\widehat{\Sigma}_i^{\mathfrak{s}}$ across all epochs. Owing to the maximum entropy principle, SSA uses the Kullback-Leibler (KL) divergence between Gaussian distributions to compare the epoch distributions up to their second moment. The matrix $\widehat{B}$ is thus found by minimizing the following cost function:

$$\mathcal{L}\left(\widehat{B}\right) = \sum_{i=1}^{N} D_{\mathrm{KL}}\left[\mathcal{N}\left(\widehat{\mu}_i^{\mathfrak{s}}, \widehat{\Sigma}_i^{\mathfrak{s}}\right) \ || \ \mathcal{N}(0, \mathbb{I})\right] = -\sum_{i=1}^{N}\left(\log\det\widehat{\Sigma}_i^{\mathfrak{s}} + \widehat{\mu}_i^{\mathfrak{s}\top}\widehat{\mu}_i^{\mathfrak{s}}\right), \tag{4}$$

where D_{KL} is the KL divergence and $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate Gaussian distribution with mean μ and covariance Σ.

2.2. Symmetrized matrix divergence

SSA and its variants use in their cost function the KL divergence between Gaussian distributions. In what follows we assume that these distributions have a zero mean. Under this assumption the KL divergence is a Bregman matrix divergence (Banerjee et al., 2005) between covariance matrices. The family of Bregman matrix divergences is generally defined as

$$D_\Phi(X, Y) = \Phi(X) - \Phi(Y) - \mathrm{tr}\left(\left(\nabla\Phi(Y)\right)^\top (X - Y)\right). \tag{5}$$

D_{KL} is obtained for $\Phi(X) = -\log\det X$, so it is often called the log-determinant divergence (Kulis et al., 2009).

Bregman matrix divergences are useful in machine learning and have a number of useful properties (Kulis et al., 2009; Banerjee et al., 2005) such as linearity and convexity in the first argument (and, in the case of the KL divergence, also in the second). However, as can be seen from their definition, they are asymmetric and do not satisfy the triangle equality. In particular, we have that $D_{\mathrm{KL}}(X \,||\, Y) \neq D_{\mathrm{KL}}(Y \,||\, X)$ for two arbitrary matrices $X \neq Y$. Subsequently, symmetrized versions of the Bregman matrix divergence, namely Jensen-Bregman divergences, have been studied in recent years (Nielsen and Nock, 2009; Banerjee et al., 2009). For the KL divergence (for zero mean distributions), this gives

$$\begin{aligned} D_{\mathrm{JBLD}}(X, Y) &= \frac{1}{2}\left[D_{\mathrm{KL}}\left(X \,||\, \frac{1}{2}(X+Y)\right) + D_{\mathrm{KL}}\left(Y \,||\, \frac{1}{2}(X+Y)\right)\right] \\ &= \log\det\left(\frac{1}{2}(X+Y)\right) - \frac{1}{2}\log\det(XY) \end{aligned} \tag{6}$$

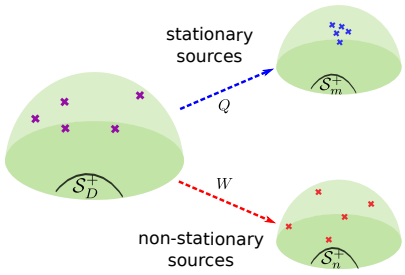

Figure 1: An illustration of our covariance-based approach. A set of input covariance matrices $\{\Sigma_i \in \mathcal{S}_D^+\}$ are made up of a mixture (purple) of stationary (blue) and non-stationary (red) parts. Due to non-stationarities, in the original space the matrices are spread out. The matrices are mapped to two lower dimensional spaces, the $\mathfrak{s}-$ (stationary) and $\mathfrak{n}-$ (non-stationary) spaces, via the matrices Q and W, respectively. In the $\mathfrak{s}$-space the matrices are now more localized compared to the $\mathfrak{n}$-space in which they have a high variance.

and is called the Jensen-Bregman log-determinant (JBLD) divergence (Cherian et al., 2011).

The JBLD has many favorable properties (see Sra (2011)), primarily that its square root comprises a metric on the SPD matrix manifold. Moreover, in Sra (2011) it has been shown that it is a close approximation to the *affine invariant Riemannian metric* (AIRM) (Bhatia, 2009) and shares many of its mathematical properties. The practical properties of both metrics, specifically their invariance properties, will be discussed in Section 2.5. In the context of SPD matrices, the JBLD is referred to as the symmetric Stein divergence or the (square of the) log-determinant metric (Sra, 2011). In the sequel we adopt the notation $\delta_{\mathrm{s}}(X, Y)$.

Motivated by the above, we reformulate SSA using a cost function based on δ_{s} (cf. Eqs. (3) and (4)):

$$\mathcal{L}(\widehat{B}) = \sum_{i=1}^{N} \delta_{\mathrm{s}}^2(\widehat{\Sigma}_i^{\mathfrak{s}}, \mathbb{I}) = \sum_{i=1}^{N} \delta_{\mathrm{s}}^2(Q^\top \tilde{\Sigma}_i Q, \mathbb{I}) = \sum_{i=1}^{N} \left[\log\det\left(\frac{1}{2}(\widehat{\Sigma}_i^{\mathfrak{s}} + \mathbb{I})\right) - \frac{1}{2}\log\det(\widehat{\Sigma}_i^{\mathfrak{s}}) \right], \quad (7)$$

where $Q = \left(\mathbb{I}_D^m \widehat{B}\right)^\top$ and $\tilde{\Sigma}_i = Z\Sigma_i Z^\top$ are the matrices whitened with $Z = \bar{\Sigma}^{-1/2}$, $\bar{\Sigma} = \underset{\Sigma \in \mathcal{S}_D^+}{\operatorname{argmin}} \sum_{i=1} \delta_{\mathrm{s}}^2(\Sigma_i, \Sigma)$, the mean of Σ_i w.r.t. δ_{s}. $\mathcal{S}_D^+$ denotes the set of all $D \times D$ SPD matrices.

2.3. A geometric interpretation

By replacing the cost function with one based on the symmetrized divergence we gain not only the benefits of the symmetric divergence, but also new insight into the problem of SSA. Note that in Eq. (7) the problem is ultimately framed in terms of distances between SPD

matrices. This suggests adopting a geometric perspective, whereby the notion of stationarity is captured by the dispersion of the matrices Σ_i. In this view, the assumption that the covariance matrices of stationary signals do not vary much between epochs translates to them having small distances between them.

An illustration of this idea is presented in Fig. 1. In this figure, the matrices Σ_i are seen as points on the SPD matrix manifold $\mathcal{S}_D^+$. The goal of our method is to find transformations Q and W that map the matrices onto two separate manifolds of lower dimension - the stationary and non-stationary space, respectively. The matrices in the stationary space will exhibit small variation, while the non-stationarities will be captured in the non-stationary space where the variation of the matrices will be greater. The transformations Q and W may be chosen to be orthogonal to each other, producing well separated $\mathfrak{s}$- and $\mathfrak{n}$-spaces. That is, $W \in \mathcal{Q}^\perp$ for $\mathcal{Q} = \operatorname{span} Q$ (likewise $Q \in \mathcal{W}^\perp$), where $\perp$ denotes the orthogonal complement.

Put formally, our objective is to find a rank-m transformation matrix $Q \in \mathbb{R}^{D\times m}$ which maps $\Sigma_i \in \mathcal{S}_D^+$ to $\widehat{\Sigma}_i^{\mathfrak{s}} \in \mathcal{S}_m^+$ for $m < D$ such that the log-determinant distance between the compressed whitened matrices $\widehat{\Sigma}_i^{\mathfrak{s}} = Q^\top \tilde{\Sigma}_i Q$ and their mean, which for the whitened matrices is $\mathbb{I}$, is minimized. Note that the space spanned by the columns of Q is of importance, and not the specific columns themselves. So, we may optimize Q over the Grassmann manifold (Edelman et al., 1998), $\mathcal{G} = \left\{\operatorname{span}(Q) \;:\; Q \in \mathbb{R}^{D\times m},\; Q^\top Q = \mathbb{I}\right\}$, the set of all m-dimensional linear subspaces of $\mathbb{R}^{D\times D}$.

In practice, for optimization we employ a Riemannian trust-regions method described in Absil et al. (2009) and implemented efficiently in Boumal et al. (2014). We add that, once the problem is framed in geometric terms, the previous zero-mean assumption becomes unnecessary. This is because the SPD (covariance) matrices, which are the focus of our work, encode the second moment of the data distribution.

2.4. A generic geometric formulation

Given the strong relation between the log-determinant metric and the AIRM (Bhatia, 2009), a natural progression is to incorporate the AIRM into the cost function. To understand why it would be beneficial to use the AIRM it is necessary first to briefly discuss the geometry of the SPD matrix manifold.

When equipped with the Frobenius inner product $\langle A, B\rangle_{\mathcal{F}} = \operatorname{tr}(A^\top B)$, the set $\mathcal{S}_n^+$ of SPD matrices of size $n \times n$, belongs to a Euclidean space. In this case, similarity between SPD matrices can be measured simply by using the Euclidean distance derived from the Euclidean norm. This is readily seen in the following example for 2×2 SPD matrices. A matrix $A \in \mathcal{S}_2^+$ can be written as $A = \left[\begin{smallmatrix} a & c \\ c & b \end{smallmatrix}\right]$ with $ab - c^2 > 0$, $a > 0$ and $b > 0$. Then matrices in $\mathcal{S}_2^+$ can be represented as points in $\mathbb{R}^3$ and the constraints can be plotted as a convex cone whose interior is populated by the SPD matrices (see Fig. 2). In this representation, the Euclidean geometry of symmetric matrices implies that distances are computed along straight lines.

Despite its simplicity, the Euclidean geometry has several drawbacks and is not always well suited for SPD matrices (Fletcher et al., 2004; Arsigny et al., 2007; Sommer et al., 2010). For example, due to an artifact referred to as the *swelling effect* (Arsigny et al., 2007), for a task as simple as averaging two matrices, it may occur that the determinant

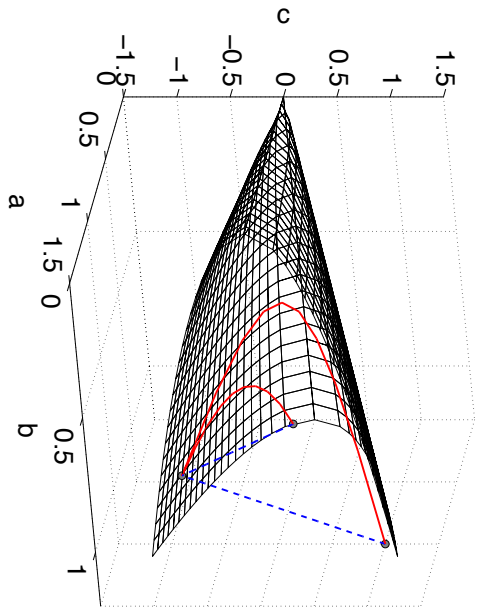

Figure 2: Comparison between Euclidean (blue straight dashed lines) and Riemannian (red curved solid lines) distances measured between points of the space $\mathcal{S}_2^+$.

of the average is larger than any of the two matrices. Another drawback, illustrated in Fig. 2 and documented by Fletcher et al. (2004), is the fact that this geometry forms a non-complete space. Hence, in this Euclidean space interpolation between SPD matrices is possible, but extrapolation may produce indefinite matrices, leading to uninterpretable solutions.

An efficient alternative which addresses these issues is to consider the space of SPD matrices as a curved space, namely a Riemannian manifold. Of the possible Riemannian distances, the AIRM, due to its favorable mathematical properties, is widely used in many applications (see, for example Fletcher et al. (2004); Pennec et al. (2006)). It is defined for any $X, Y \in \mathcal{S}_D^+$ as:

$$\delta_{\mathrm{r}}^2(X,Y) = \left\| \log\left(X^{-1/2} Y X^{-1/2}\right) \right\|_{\mathbb{F}}^2, \tag{8}$$

where $\log(\cdot)$ is the matrix logarithm and $\|X\|_{\mathbb{F}}^2 = \mathrm{tr}\left(X^\top X\right)$ is the Frobenius norm.

In the curved space, the geodesics between matrices obtained by the AIRM are computed on curved lines as illustrated in Fig. 2 for the space $\mathcal{S}_2^+$. Symmetric matrices with null and infinite eigenvalues (i.e., those which lie on the boundary of the convex cone, but not in it) are both at an infinite distance from any SPD matrix on the manifold (*within* the cone).

So, let us now consider a cost function of the same form defined w.r.t. the AIRM.[2] A general expression for our gaSSA is then:

$$\widehat{Q} = \underset{Q \in \mathcal{G}(D,m)}{\operatorname{argmin}} \sum_i \delta^2 \left(Q^\top \tilde{\Sigma}_i Q, \mathbb{I} \right), \tag{9}$$

where $\tilde{\Sigma}_i = Z \Sigma_i Z^\top$ are the matrices whitened with $Z = \bar{\Sigma}^{-1/2}$ and $\bar{\Sigma} = \underset{\Sigma \in \mathcal{S}_D^+}{\operatorname{argmin}} \sum_{i=1} \delta^2 (\Sigma_i, \Sigma)$ is the matrix mean w.r.t. δ. In the next section we will show that the need for matrix whitening can be alleviated.

We note that this cost function is similar in spirit to an unsupervised version of the one in Harandi et al. (2014). For reference, the Euclidean gradient w.r.t. Q of the cost function, used for the optimization, can be found in Horev et al. (2015).

2.5. Symmetries and invariance properties

We now discuss the symmetries of our optimization problem and the invariance properties of our chosen metrics. These properties will enable us to significantly simplify our problem and eliminate the matrix whitening step. For brevity we will state the results in terms of the AIRM, but the same holds true for the log-determinant metric.

Our key observation stems from the fact that δ_{r} and δ_{s} are invariant to congruent transformations of the form $X \mapsto P^\top X P$ for $P \in GL_D(\mathbb{R})$ (Bhatia, 2009).[3] So, we have

$$\delta_{\mathrm{r}}^2 (X, Y) = \delta_{\mathrm{r}}^2 \left(P^\top X P, P^\top Y P \right). \tag{10}$$

for $X, Y \in \mathcal{S}_D^+$ and a real-valued invertible matrix P. This is a crucial point since both the whitening matrix Z and the *mixing matrix* A act on the covariance matrices in this way.

Proposition 1 *Let $\mathbf{\Lambda} = \{\Lambda_i\}_{i=1}^n$ for $\Lambda_i \in \mathcal{S}_n^+$ be a set of SPD matrices of size $n \times n$ and let $\Sigma_i = A \Lambda_i A^\top$ for some real invertible matrix A. Denoting the Riemannian mean (the mean w.r.t. δ_{r}) of $\mathbf{\Lambda}$ by $\bar{\Lambda}$, the Riemannian mean $\bar{\Sigma}$ of the set $\mathbf{\Sigma} = \{\Sigma_i\}_{i=1}^n$ is given by $\bar{\Sigma} = A \bar{\Lambda} A^\top$.*

Proof The Riemannian mean of the set $\mathbf{\Lambda}$ is defined as $\bar{\Lambda} = \underset{\Lambda \in \mathcal{S}_n^+}{\operatorname{argmin}} \sum_i \delta_{\mathrm{r}}^2 (\Lambda_i, \Lambda)$. Using the congruence invariance (Eq. (10)) we have $\delta_{\mathrm{r}}^2 (\Lambda_i, \Lambda) = \delta_{\mathrm{r}}^2 \left(\Sigma_i, A \Lambda A^\top \right)$, so $A \bar{\Lambda} A^\top$ is the minimizer of $\sum_i \delta_{\mathrm{r}}^2 (\Sigma_i, \Sigma)$ and the result follows. ■

Using the above and by simple manipulation we obtain the following equivalence relations.

Corollary 2 *The following expressions are equivalent:*

$$\delta_{\mathrm{r}}^2 \left(\tilde{\Sigma}_i, \mathbb{I} \right) = \delta_{\mathrm{r}}^2 \left(\Sigma_i, \bar{\Sigma} \right) = \delta_{\mathrm{r}}^2 \left(\Gamma_i, \bar{\Gamma} \right), \tag{11}$$

where Γ_i is the covariance matrix of the unmixed *sources in the i-th epoch.*

2. Other metrics such as the Euclidean metric or the log-Euclidean metric (Arsigny et al., 2007) may also be used.
3. This also holds for complex matrices $P \in GL_D(\mathbb{C})$ where the matrix transpose is replaced by the conjugate transpose $P^H = \overline{P}^\top$.

We have shown that both the whitening operation and the mixing matrix A do not affect the distance between the covariance matrices of the original unmixed signals. We can then re-write our optimization problem as

$$\begin{aligned}\widehat{Q} &= \underset{Q\in\mathcal{G}(D,m)}{\operatorname{argmin}} \sum_i \delta^2\left(Q^\top \tilde{\Sigma}_i Q, \mathbb{I}\right) = \underset{Q\in\mathcal{G}(D,m)}{\operatorname{argmin}} \sum_i \delta^2\left(Q^\top A\Gamma_i A^\top Q, Q^\top A\bar{\Gamma}A^\top Q\right) \\ &= \underset{Q'\in\mathcal{G}(D,m)}{\operatorname{argmin}} \sum_i \delta^2\left(Q'^\top \Gamma_i Q', Q'^\top \bar{\Gamma} Q'\right). \end{aligned} \tag{12}$$

One may remark that $A^\top Q$ no longer has orthonormal columns and so does not belong to the Grassmann manifold. While this is indeed true, the final transition is due to the observation that the solution to our problem is not unique. Our interest is in recovering the stationary subspace and not the exact sources themselves, so the solution is invariant to any transformation (e.g., subspace scaling and rotation) acting within each of the $\mathfrak{s}$- and $\mathfrak{n}$-spaces separately. Furthermore, we have chosen $W \in \mathcal{Q}^\perp$, and so the $\mathfrak{s}$-space is orthogonal to the $\mathfrak{n}$-space. Now, choosing orthogonal bases within each of the subspaces we may restrict ourselves to *orthogonal* mixing matrices A and find a transformation Q' which lies in the Grassman manifold.

The final result is quite remarkable. First, it shows that our problem is essentially agnostic to the mixing matrix. Secondly, it eliminates the need to whiten the matrices. This is useful when the covariance matrices are poorly estimated and data whitening introduces additional error, for example, when the epochs are short compared to the number of signals or when data is corrupted by noise.

In conclusion, we have two variations of gaSSA given in the first and last terms of Eq. (12). The difference between the two is whether or not the input covariance matrices are whitened. Our analysis shows that whitening does not improve performance, and may in fact lead to a degradation of the results in certain cases. So, we claim that it is in general preferable not to whiten the matrices. In terms of the chosen metric, we do not expect a significant difference when using δ_{r} vs. δ_{s}. In the following section we will present experimental evidence to support these claims.

3. Experimental results

In this section we present experimental results on synthetic data and data taken from real BCI experiments. We compare the performance of gaSSA to the existing SSA and investigate the effects of matrix whitening and the choice of metric.

3.1. Toy data

For our first experiment we generated data following the SSA model as a mixture of stationary and non-stationary sources. To generate non-stationarity in the data we used a slightly modified version of the scheme provided in the SSA toolbox (Müller et al., 2011) and detailed in its user manual. Here we bring only a brief description:

The elements of the mixing matrix A are chosen uniformly from the range $[-0.5, 0.5]$ and its columns are normalized to 1. The distribution of the $\mathfrak{s}$-sources is constant over all epochs, namely $s^{\mathfrak{s}}(t) \sim \mathcal{N}(0, \Lambda^{\mathfrak{s}})$. In the SSA toolbox, $\Lambda^{\mathfrak{s}}$ is taken to be the identity matrix,

however we choose $\Lambda^{\mathfrak{s}}$ to be a random matrix of the form $\Lambda^{\mathfrak{s}} = B\Gamma B^\top$ for an orthogonal matrix B and diagonal matrix Γ.

The $\mathfrak{n}$-sources are correlated with the $\mathfrak{s}$-sources, and for the i-th epoch $\tau_i = [t_0(i), \ldots, t_0(i) + T]$ they are given by $s^{\mathfrak{n}}(t) = C_i s^{\mathfrak{s}}(t) + Y^{\mathfrak{n}}(t)$ for $t \in \tau_i$, where $C_i \in \mathbb{R}^{(D-m)\times m}$ and $Y^{\mathfrak{n}}(t) \sim \mathcal{N}(\mu_i, \Lambda_i^{\mathfrak{n}})$. The covariance matrices $\Lambda_i^{\mathfrak{n}}$ are generated for each epoch in the same way as $\Lambda^{\mathfrak{s}}$.

So, the covariance matrix of the (unmixed) sources in the i-th epoch may be written as

$$\Lambda_i = \operatorname{cov}\left(\begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}\right) = \begin{bmatrix} \Lambda^{\mathfrak{s}} & (C_i\Lambda^{\mathfrak{s}})^\top \\ C_i\Lambda^{\mathfrak{s}} & C_i\Lambda^{\mathfrak{s}}C_i^\top + \Lambda_i^{\mathfrak{n}} \end{bmatrix}. \tag{13}$$

Using the data generated by the scheme above we compared the performance of gaSSA to that of SSA. As a performance measure we used the distance between the estimated $\mathfrak{n}$-space $\widehat{A}^{\mathfrak{n}}$ and the true $\mathfrak{n}$-space $A^{\mathfrak{n}}$. This is owing to the fact that, as discussed in von Bünau et al. (2009b), the $\mathfrak{n}$-space and $\mathfrak{s}$-sources are identifiable, while the $\mathfrak{s}$-space and $\mathfrak{n}$-sources are not. To illustrate this, note that to be stationary, the $\mathfrak{s}$-sources must consist strictly of stationary sources, while the $\mathfrak{n}$-sources will remain non-stationary even if they are mixed with stationary signals. The distance between sub-spaces is computed using $\delta_{\mathcal{G}}$, the metric on the Grassmann manifold (Absil et al., 2004). Shortly, this metric is based on the principal angles between the two spaces.

We generated 50 epochs of length $T = 250$ for several values of D and m. For each pair (D, m) we conducted the experiment 25 times. At each iteration the optimization was restarted 5 times with different initial guesses and the transformation matrix which obtained the lowest cost was selected. The results are summarized in Table 1.

Our gaSSA method obtained results roughly twice as good as those of SSA for all iterations. The improved performance of our method can be attributed to three factors: optimization over the Grassmann manifold, the lack of matrix whitening and the geometric cost function itself. Although the optimization is carried out over different manifolds in SSA (Grassmann manifold) and gaSSA (rotation manifold), both manifolds capture the invariance properties of the solution. Since both problems are non-convex, it is difficult to assess which optimization problem allows for better minimization of its corresponding cost function. Our analysis in Section 2.5 shows that in general the scheme with and without matrix whitening should produce the same solution. Indeed, we see that the error in the ns-subspace estimation is essentially identical for the two schemes. This suggests that it is the new geometric objective function itself which improves the estimation of the stationary subspace.

3.2. Brain-computer interface

Next, we applied our method to data taken from the BCI competition IV dataset II. This dataset contains motor imagery (MI) EEG signals affected by eye movement artifacts. It was collected in a multi-class setting, with the subjects performing more than 2 different MI tasks. However, as in Lotte and Guan (2011), we evaluate our algorithms on two-class problems by selecting only signals of left- and right-hand MI trials.

We applied the same pre-processing as described in Lotte and Guan (2011). EEG signals were band-pass filtered in $8 - 30$ Hz, using a 5^{th} order Butterworth filter. For each trial, we

Table 1: Average error in subspace estimation for various numbers of stationary and non-stationary signals. (w) and (nw) signify that matrix whitening was / was not performed. The standard deviation of the results appears in parentheses.

D	m	gaSSA δ_r (w)	gaSSA δ_r (nw)	gaSSA δ_s (w)	gaSSA δ_s (nw)	SSA
19	7	0.0018 (1e-4)	0.0018 (1e-4)	0.0018 (1e-4)	0.0018 (1e-4)	0.004 (1e-4)
19	5	0.0022 (1e-4)	0.0022 (1e-4)	0.0022 (1e-4)	0.0022 (1e-4)	0.0123 (2e-4)
13	7	0.001 (1e-4)	0.001 (1e-4)	0.001 (1e-4)	0.001(1e-4)	0.0042 (1.5e-4)
13	5	0.0025 (1e-4)	0.0025 (1e-4)	0.0025 (1e-4)	0.0025 (1e-4)	0.0045 (2e-4)

extracted features from the time segment located from 0.5s to 2.5s after the cue instructing the subject to perform MI.

The data was initially divided into two parts: a training data set and a test data set. Similarly to von Bünau et al. (2009b) the first 20% of the test trials were set aside for adaptation. The aim of the adaptation part is to mitigate any non-stationarities between the test and the training session. We then learned the $\mathfrak{s}$-space in an *unsupervised* manner over the training and adaptation part. As before, our method was reinitialized 5 times and the transformation attaining the lowest cost was chosen.

The performance was measured by means of the classification rate on the test set. We used the following naive classifier, referred to as *minimum distance to the mean* (MDM) in Barachant et al. (2012): Using the labels of the training set, we compute the mean (in the $\mathfrak{s}$-space) for each of the two classes. Then, we classify the compressed covariance matrices in the test set according to their distance to the class means; each matrix is assigned the class to which it is closer.

The original data is comprised of 22 signals. Since the true number of stationary signals is unknown, we repeated the experiment for several values in the range $m \in [10, 18]$. The results for the nine subjects in the dataset are summarized in Table 2.

The results show that our method outperforms SSA for most subjects. As m grows and there are assumed to be less non-stationary components, the problem of SSA becomes simpler. Our method can better identify the few key elements that contribute most to the stationary subspace. As we search for more components, their significance is diminished and the gap in performance between SSA and gaSSA decreases. In terms of the metric, we see that δ_r and δ_s perform roughly the same. The schemes that performed matrix whitening generally achieved lower accuracy than those which did not. In this complex setting, more accurate estimation of the mixing matrix does not guarantee better classification.

Initially these results may seem contradictory to the analysis of section 2.5, however the results of the whitening/non-whitening versions differ for two reasons. The first is the non-convexity of our problem and its dependency on initialization. In our experiments we initialize both versions with the same matrix. However, due to the action of the whitening matrix the subspace spanned by the initial Q is effectively rotated, yielding a different initial subspace (the reasoning here is similar to that of the final transition in Eq. (12)). Secondly, we note that the test set is whitened with its own Riemannian mean. After whitening

Table 2: Classification accuracy for (from top to bottom) $m = 10, 12, 14, 16, 18$ $\mathfrak{s}$-sources. Best results are highlighted in boldface. (w) and (nw) signify that matrix whitening was / was not performed.

subject #	1	2	3	4	5	6	7	8	9	avg
gaSSA δ_{r} (w)	48.96	55.65	63.48	58.22	**60.69**	56.26	68.09	71.30	48.70	59.04
gaSSA δ_{r} (nw)	73.91	59.13	**91.3**	**69.3**	57.67	**66.43**	58.70	**93.91**	**86.09**	**72.94**
gaSSA δ_{s} (w)	49.83	55.65	63.57	61.43	58.9	55.7	**68.17**	72.17	47.34	59.2
gaSSA δ_{s} (nw)	**74.13**	**60**	**91.3**	68.52	57.03	63.96	59.48	**93.91**	85.99	72.7
SSA	47.74	57.16	60	57.61	56.02	54.17	67.3	75.89	53.82	58.86
gaSSA δ_{r} (w)	59.26	56.26	71.17	62.61	58.26	57.39	**70.68**	72.70	73.04	64.6
gaSSA δ_{r} (nw)	**73.96**	57.39	**93.83**	68.7	58.26	**66.09**	62.61	**94.26**	**86.09**	**73.46**
gaSSA δ_{s} (w)	59.13	56.39	68.7	60.87	58.09	57.39	70.43	72.81	71.3	63.9
gaSSA δ_{s} (nw)	73.04	**58.26**	92.17	**69.57**	58.3	64.35	62.61	93.28	**86.09**	73.07
SSA	57.96	52.91	61.48	61.04	**62.91**	53.78	65.28	74.78	66.16	61.81
gaSSA δ_{r} (w)	65.02	53.91	72.35	65.22	**62.61**	59.13	72.93	77.39	77.39	67.33
gaSSA δ_{r} (nw)	**74.4**	**58.26**	88.74	67.83	60	**65.22**	68.91	**94.78**	**87.83**	**74**
gaSSA δ_{s} (w)	65.02	53.91	72.35	65.22	**62.61**	58.87	72.83	77.39	76.52	67.19
gaSSA δ_{s} (nw)	**74.4**	**58.26**	**88.78**	67.83	60	63.83	68.15	93.91	**87.83**	73.67
SSA	67.05	54.72	66.7	**68.7**	56.87	55.57	**73.48**	75.65	77.39	66.24
gaSSA δ_{r} (w)	68.17	60.78	86.17	70.43	62.61	58.26	74.78	84.35	82.46	72
gaSSA δ_{r} (nw)	**78.43**	60.87	**90.26**	**73.04**	59.3	67.83	67.83	**94.78**	**88.77**	**75.68**
gaSSA δ_{s} (w)	66.78	**61.65**	86.78	70.43	**62.74**	56.52	74.78	83.48	81.74	71.65
gaSSA δ_{s} (nw)	76.87	60.3	90.09	71.3	60	**68.7**	67.83	**94.78**	88.7	75.4
SSA	74.83	56.3	76.52	68.7	62.3	63.35	**75.59**	85.22	79.06	71.32
gaSSA δ_{r} (w)	73.04	60.13	86.96	**72.78**	61.74	62.61	72.17	86.09	85.22	73.41
gaSSA δ_{r} (nw)	**85.22**	60.96	**90.43**	72.35	61.74	**69.57**	71.3	**94.78**	**88.7**	**77.23**
gaSSA δ_{s} (w)	70.43	60.83	86.96	**72.78**	61.74	61.74	73.04	86.09	85.22	73.2
gaSSA δ_{s} (nw)	84.35	**61.09**	**90.43**	72.35	61.74	**69.57**	68.7	**94.78**	**88.7**	76.86
SSA	71.3	60.61	78.09	70.3	**65.39**	65.22	**73.79**	84.35	80.87	72.21

the position of the test set relative to the training is different than when the sets are not whitened. This changes the results of our classifier, but not the subspace estimation.

4. Conclusion

We presented a covariance-based method for unsupervised stationary subspace analysis. The problem was phrased in terms of the distance between matrices.Owing to the symmetries of the problem and the invariance properties of the metrics, we derived useful equivalence relations that showed that it is not necessary to whiten the input covariance matrices. Experiments on both synthetic and BCI data supported our theoretical analysis and showed that our method outperforms SSA.

In the future we wish to tackle the challenges stemming from to different types of non-stationarity, occurring both within sessions (intra-session) and between them (inter-session),

and due to the introduction of class-wise variation which must not be discarded as non-stationarity in classification tasks. A promising application is change point detection, where more accurate estimation of the $\mathfrak{n}$-space may lead to better detection of change points.

Acknowledgments

During this work, IH was supported in part by a MEXT scholarship. IH and MS would like to acknowledge the support of KAKENHI 25700022.

References

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80(2):199–220, 2004.

Pierre-Antoine Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2009.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

Arindam Banerjee, Daniel Boley, and Sreangsu Acharyya. Symmetrized Bregman divergences and metrics. In *The Learning Workshop*, volume 2. Citeseer, 2009.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomputing*, 112: 172–178, 2013.

Rajendra Bhatia. *Positive Definite Matrices*. Princeton University Press, 2009.

Nicolas Boumal, Bamdev Mishra, Pierre-Antoine Absil, and Rodolphe Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014. URL www.manopt.org.

Anoop Cherian and Suvrit Sra. Riemannian sparse coding for positive definite matrices. In *European Conference on Computer Vision (ECCV)*, pages 299–314, 2014.

Anoop Cherian, Suvrit Sra, Arindam Banerjee, and Nikolaos Papanikolopoulos. Efficient similarity search for covariance matrices via the Jensen-Bregman log-det divergence. In *2011 IEEE International Conference on Computer Vision (ICCV)*, pages 2399–2406. IEEE, 2011.

Israel Cohen and Baruch Berdugo. Speech enhancement for non-stationary noise environments. *Signal processing*, 81(11):2403–2418, 2001.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Thomas P. Fletcher, Conglin Lu, Stephen M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging*, 23(8): 995–1005, 2004.

Mehrtash Harandi, Mathieu Salzmann, and Richard Hartley. From manifold to manifold: geometry-aware dimensionality reduction for spd matrices. In *European Conference on Computer Vision (ECCV)*, pages 17–32, 2014.

Inbal Horev, Florian Yger, and Masashi Sugiyama. Intrinsic PCA for SPD matrices. In *Asian Conference on Machine Learning (ACML)*, 2015.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*, volume 46. John Wiley & Sons, 2004.

Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*, volume 7. Cambridge University Press, 2004.

Brian Kulis, Mátyás A. Sustik, and Inderjit S. Dhillon. Low-rank kernel learning with Bregman matrix divergences. *The Journal of Machine Learning Research*, 10:341–376, 2009.

Matt J. Kusner, Nicholas I. Kolkin, Stephen Tyree, and Kilian Q. Weinberger. Stochastic covariance compression. *arXiv preprint arXiv:1412.1740*, 2014.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Fabien Lotte and Cuntai Guan. Regularizing common spatial patterns to improve bci designs: unified theory and new algorithms. *IEEE Transactions on Biomedical Engineering*, 58(2):355–362, 2011.

Jan Saputra Müller, Paul von Bünau, Frank C. Meinecke, Franz J. Király, and Klaus-Robert Müller. The stationary subspace analysis toolbox. *The Journal of Machine Learning Research*, 12:3065–3069, 2011.

Frank Nielsen and Richard Nock. Sided and symmetrized Bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2882–2904, 2009.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

Wojciech Samek, Klaus-Robert Müller, Motoaki Kawanabe, and Carmen Vidaurre. Brain-computer interfacing in discriminative and stationary subspaces. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2873–2876. IEEE, 2012.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Stefan Sommer, François Lauze, Søren Hauberg, and Mads Nielsen. Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In *European Conference on Computer Vision*, pages 43–56, 2010.

Suvrit Sra. Positive definite matrices and the s-divergence. *arXiv preprint arXiv:1110.1773*, 2011.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments*. The MIT Press, 2012.

Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision*, pages 589–600, 2006.

Paul von Bünau, Frank C. Meinecke, Franz C. Király, and Klaus-Robert Müller. Finding stationary subspaces in multivariate time series. *Physical Review Letters*, 103(21):214101, 2009a.

Paul von Bünau, Frank C. Meinecke, and Klaus-Robert Müller. Stationary subspace analysis. In *Independent Component Analysis and Signal Separation*, pages 1–8. Springer, 2009b.

Wojciech Wojcikiewicz, Carmen Vidaurre, and Motoaki Kawanabe. Stationary common spatial patterns: towards robust classification of non-stationary eeg signals. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 577–580. IEEE, 2011.