Wasserstein Stationary Subspace Analysis

Stephan Kaltenstadler, Shinichi Nakajima, Klaus-Robert Müller*, *Member, IEEE*,
and Wojciech Samek*, *Member, IEEE*

***Abstract*—Learning under non-stationarity can be achieved by decomposing the data into a subspace that is stationary and a non-stationary one (stationary subspace analysis (SSA)). While SSA has been used in various applications, its robustness and computational efficiency has limits due to the difficulty in optimizing the Kullback-Leibler divergence based objective. In this paper we contribute by extending SSA twofold: we propose SSA with (a) higher numerical efficiency by defining analytical SSA variants and (b) higher robustness by utilizing the Wasserstein-2 distance (Wasserstein SSA). We show the usefulness of our novel algorithms for toy data demonstrating their mathematical properties and for real-world data (1) allowing better segmentation of time series and (2) brain-computer interfacing, where the Wasserstein-based measure of non-stationarity is used for spatial filter regularization and gives rise to higher decoding performance.**

***Index Terms*—Subspace learning, stationary subspace analysis, divergence methods, optimal transport, covariance metrics.**

I. Introduction

SUBSPACE methods are one of the most common basic tools in many research fields, including machine learning, signal processing, image processing, computer vision, natural language processing, e-commerce, and bioinformatics. Since the available data size, dimension and multi-modality has grown explosively, the importance of subspace methods has increased. With that the robustness against outliers became an essential research topic, since manual outlier rejection is practically impossible on large-scale datasets. Different approaches and their robust variants have been developed to find suitable subspaces for special problems (e.g., [1]).

Subspace methods in general decompose high dimensional data X into a low-dimensional source S and a projection A such that

$$X \approx AS.$$

Various additional assumptions on S, which make the decomposition unique, define different subspace methods. Principal component analysis (PCA) chooses a low dimensional subspace that retains the variance as much as possible. This can be achieved by finding the leading eigendirections of the covariance matrix. Independent component analysis (ICA) [2] assumes independence of the source signals and minimizes the dependence between the components. A popular independence criterion to be maximized is the negentropy between the sources, which corresponds to maximizing the distance of the estimated source distributions from a Gaussian distribution, as can be seen for example in [3].

This paper focuses on stationary subspace analysis (SSA), proposed by Bünau et al. [4], which is a projection algorithm splitting the multivariate data stream into stationary and non-stationary parts. Stationarity of the data distribution is a common assumption in many applications in machine learning, since it assures the convergence of most estimators. However, in practice this assumption is rarely satisfied [5]. SSA assumes that the data is generated from a mixture of sources, similarly to ICA, and the sources consist of stationary and non-stationary components. Thus, this method employs changes in the data time-structure to find a decomposition, which maximizes/minimizes *stationarity* in the data.

Stationarity is evaluated by measuring the distance between distributions in each epoch of the time series. SSA approximates the underlying distributions with Gaussian distributions, and their distance is measured by the Kullback-Leibler (KL) divergence. But since the Kullback-Leibler divergence is not a proper metric (see [6] for more information about divergences), it lacks symmetries and invariance properties. Horev et al. [7] proposed a geometry-aware variant of SSA by adopting the affine invariant Riemann metric.

In this paper, we study a particular choice for the distance measure. Namely, we propose Wasserstein SSA (WaSSA), where the distance between distributions is measured by the Wasserstein-2 distance [8]. The advantage of this choice is twofold:

1) *Computational efficiency*: WaSSA can be carried out by eigendecomposition after a minor iterative optimization which typically converges in 5-7 steps. This gives a significant computational advantage against most of the existing methods, which are solved by gradient descent style algorithms. We also propose an approximate variant of WaSSA using the matrix-root distance, which further improves the computational efficiency.
2) *Robustness against outliers*: Under some assumption, the (square) Wasserstein distance is written as a sum of L2-distance between the means and the matrix-root distance between the covariances. Since the matrix-root in general downweights the contribution from peaky data points, WaSSA is expected to be robust against outliers.

In this work we discuss some mathematical properties of the Wasserstein distance and the geometry induced by it and a related metric, the matrix-root distance. The resulting algorithm is solvable by an eigendecomposition combined with

* Corresponding authors
Manuscript received April 16, 2018.
S. Kaltenstadler and W. Samek are with Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany (e-mail: stephan.kaltenstadler@hhi.fraunhofer.de; wojciech.samek@hhi.fraunhofer.de).

S. Nakajima is with the Technische Universität Berlin, 10587 Berlin, Germany (e-mail: nakajima@tu-berlin.de).

K.-R. Müller is with the Technische Universität Berlin, 10587 Berlin, Germany, and the Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea and also with Max Planck Institute for Informatics, Saarbrücken, Germany (e-mail: klaus-robert.mueller@tu-berlin.de).

a fixed point iteration. Additionally we discuss SSA from the perspective of a covariance estimation problem, under which SSA can be considered as a PCA-like algorithm on the space of covariance matrices. An experimental evaluation shows clear advantages over the existing methods in terms of computational efficiency and accuracy. We employ a recently developed algorithm to derive the Fréchet-mean for the Wasserstein-2 distance for a set of covariance matrices and empirically investigate the relation between the Wasserstein and matrix-root distance, a topic which has recently attracted considerable theoretical attention [9]. Finally, we evaluate these new methods on EEG data by incorporating the Wasserstein objective into the stationary common spatial pattern algorithm [10], which improves the decoding performance in brain-computer interfacing (BCI) [11], [12].

This paper is organized as follows. In Section II we give a brief overview of SSA and metrics for covariance matrices. In Section III we propose the novel Wasserstein SSA algorithm and its approximate variant. In Section IV we show experimental results demonstrating advantages of Wasserstein SSA. We conclude in Section V with a brief discussion.

A. Related work

1) Wasserstein distance: The Wasserstein-2 distance for Gaussian distributions has been well studied [13], [14], [15]. This work has enjoyed recent popularity with several theoretical results e.g. by Alvarez-Esteban et al. [16] who proposed the fixed point algorithm for calculation of the Fréchet-mean, Masarotto et al. [17], who relate the Wasserstein distance to the Procrustes distance or Bhatia et al. [9] who provide some further proofs concerning these results and raise the question of the exact relation between Wasserstein distance and the matrix-root distance. Some recent applications for covariance matrices are for example by Bachoc et al. [18], who define a family of kernel matrices based on the Wasserstein distances for forecasting of Gaussian processes, or Mallasto et al. [19], who study the use of Wasserstein space on covariance operators for Gaussian processes. Further applications of Wasserstein distance are the application of Wasserstein distance without assuming a Gaussian distribution for defining objective function in neural networks, e.g. Montavon et al. [20] showed advantages of Wasserstein distance in training restricted Boltzmann machines, while Arjovsky et al. [21] used Wasserstein distance to improve the training of generative adversarial networks.

2) Stationary subspace analysis: Stationary subspace analysis is a projection algorithm splitting the multivariate data stream into a stationary and non-stationary part [4]. Several works have proposed adaptions of the SSA framework and applied them to different applications like change point detection [22], EEG data [23], geomagnetic data [24] and videos classification [25] among others. Additionally, several extensions of SSA for different objectives have been proposed: Panknin et al. [26] take into account higher moments, while Király et al. [27] discuss an algebraic solution to problems like SSA. Baktashmotlagh et al. [25] propose a supervised approach to SSA as well as a kernelized version of the algorithm. Other authors [28] investigate a information geometric interpretation of the SSA objective or define a cost function based on the Affine invariant Riemann metric [7], which is better suited to the space of covariance matrices. Hara et al. [24] derived an approximation of the SSA objective which is solvable analytically.

B. Robust methods for EEG analysis

Subspace methods enjoy popularity in EEG analysis, since they allow to address the high dimensionality and the high inherent noise level of these signals, e.g., in brain-computer interfacing (BCI). Additionally one often has to account for outliers in the data, which can occur due to movements by the participants during BCI or loose electrodes for instance. This leads to a demand for robust methods. Castells et al. [29] give an overview of PCA methods in this domain, while Lin et al. [30] as well as Chang et al. [31] adapt robust PCA (RPCA), which tries to recover a low rank representation of the data, to brain signals. The authors of [32] combine RPCA with random projection methods, while [31] use RPCA to alleviate inter-day variability in EEG data. Common spatial pattern (CSP) algorithms [33] are another class of subspace algorithms enjoying popularity in the EEG community. Therefore, various methods have been proposed to robustify CSP [34], [35], [36], [37], [38], [39] or to enforce stationarity in the filters [40], [41], [10], [42]. The usage of SSA as preprocessing step in BCI has been investigated by Bünau et al. [23], Samek et al. [43], [44] and Horev et al. [7].

II. Background

In this section we give a brief introduction of existing variants of stationary subspace analysis methods and distance metrics between covariance matrices, based on which we introduce our proposed method in Section III.

A. Stationary subspace analysis

Stationary subspace analysis separates a multivariate time series $x(t) \in \mathbb{R}^D$ into stationary and non-stationary subspaces. The generating model for the data is a linear mixture of the stationary sources $s_s(t) \in \mathbb{R}^{d_s}$ and non-stationary sources $s_n(t) \in \mathbb{R}^{d_n}$, and is given as

$$x(t) = [A_s \; A_n] \begin{pmatrix} s_s(t) \\ s_n(t) \end{pmatrix} \tag{1}$$

with $A = [A_s \; A_n]$ assumed to be invertible. By d_s we indicate the dimension of the stationary subspace, while $d_n = D - d_s$ is the dimension of the non-stationary subspace. The goal of SSA is to find the linear transformation A^{-1}, i.e., to recover the stationary and non-stationary sources $s_s(t), s_n(t)$. This is achieved by finding projections $B_s \in \mathbb{R}^{d_s \times D}$ and $B_n \in \mathbb{R}^{d_n \times D}$ such that

$$\begin{pmatrix} \hat{s}_s(t) \\ \hat{s}_n(t) \end{pmatrix} = [B_s \; B_n] x(t). \tag{2}$$

The number of stationary sources is assumed to be known. The definition of stationarity used for SSA is stationarity in the weak sense, which assumes that the first two moments of the

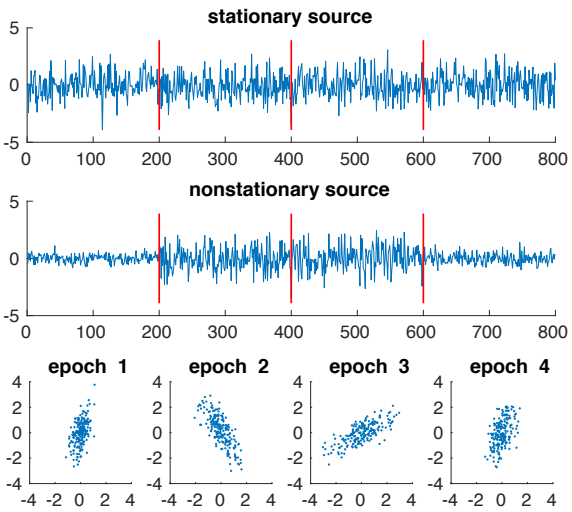

Fig. 1: Schematic for SSA: Stationary and non-stationary source with time variable covariance illustrated by the scatter plots for 4 epochs.

data do not change over time. This amounts to approximating each distribution with a Gaussian, and stationarity is evaluated based on the changes of means and covariances[1].

To exploit the temporal behaviour of the time series for this decomposition, the data is first split into several epochs (see Fig. 1), which consist of consecutive datapoints. The first and second moments of the data-epochs are computed and their variation is evaluated according to some distance measure. The chosen measure for the standard SSA algorithm is the Kullback-Leibler divergence of Gaussian distributions, i.e. SSA finds the optimal projection B_{stat} by solving the problem

$$\min_{B:BB^\top=\mathbb{I}} \sum_{e\in E} D_{KL}(\mathcal{N}(\tilde{B}\mu_e, \tilde{B}\Sigma_e\tilde{B}^\top) \parallel \mathcal{N}(0, \tilde{B}\bar{\Sigma}\tilde{B}^\top)) \quad (3)$$

with $\tilde{B} = \mathbb{I}_d B$. Here μ_e, Σ_e are the mean and covariance of the data in epoch e and $\bar{\mu}, \bar{\Sigma}$ are the mean values of mean and covariance across the whole dataset, which we assume to be centered. We further define $\mathbb{I} \in \mathbb{R}^{D\times D}$ as the D-dimensional identity matrix and $\mathbb{I}_d \in \mathbb{R}^{d\times D}$ as the projection to the first d dimensions. Note that $B \in \mathbb{R}^{D\times D}$ is a rotation matrix over which the optimization is performed, however, the objective is evaluated on the projected data, i.e., in the d-dimensional subspace. If the data is pre-whitened and writing $d_s = d$, the optimal projection for (3) can be found by solving the optimization problem

$$\begin{aligned}&\min_{B:BB^\top=\mathbb{I}} \sum_{e\in E} D_{\mathrm{KL}}(\mathcal{N}(\tilde{B}\mu_e, \tilde{B}\Sigma_e\tilde{B}^\top) \parallel \mathcal{N}(0, \mathbb{I}_{d\times d}))\\ &= \min_{B:BB^\top=\mathbb{I}} \sum_{e\in E} -\log\det \mathbb{I}_d B\Sigma_e B^\top \mathbb{I}_d^\top + \mu_e^\top B^\top \mathbb{I}_d^\top \mathbb{I}_d B\mu_e.\end{aligned} \quad (4)$$

[1]Under this model only the true stationary sources were shown to be identifiable in general, while the true non-stationary sources are unidentifiable as was discussed in Bünau et al. [4].

$\mathbb{I}_{d\times d}$ denotes the $d \times d$ identity matrix. This problem is solved by a gradient descent on the special orthogonal group SO(D), which is the subgroup of D-dimensional orthogonal matrices having determinant 1, corresponding to the group of D-dimensional rotations. Since (3), as well as (4), is a non-convex problem, standard SSA optimization is challenging and care must be exercised not to end in suboptimal solutions. As a remedy, the gradient descent algorithm is restarted multiple times (typically 5-10) from different initial points and the solution giving the lowest objective value is chosen. By symmetrizing the SSA objective, Horev et al. [7] achieve a cost function which is a proper distance on the space of semi-positive definite matrices. The original SSA problem is stated in terms of a matrix divergence, which is not a metric as it is neither symmetric nor does it satisfy the triangle inequality. Therefore they propose the use of a metric which takes into account the geometry of covariance matrices, the affine invariant Riemann metric,

$$\delta_r^2(X,Y) = ||\log(X^{-\frac{1}{2}} Y X^{-\frac{1}{2}})||_{\mathrm{fro}}^2. \quad (5)$$

This metric was shown to be invariant to congruent transformations $X \to P^\top XP$ for $P \in GL_D(\mathbb{R})$, i.e. it is invariant w.r.t. to mixing as well as whitening. They experimentally evaluated the effect of the whitening process and found it to degrade the performance.

Their proposed method, called geometry-aware SSA calculates the optimal projection B_{stat} by solving the problem

$$\min_{B:BB^\top=\mathbb{I}} \sum_{e\in E} \delta_r^2(\mathbb{I}_d B\Sigma_e B^\top \mathbb{I}_d^\top, \mathbb{I}_d B\bar{\Sigma} B^\top \mathbb{I}_d^\top). \quad (6)$$

Optimization is carried out by gradient descent on the Grassmanian manifold combined with another gradient descent algorithm to calculate the corresponding Fréchet-mean.
A different direction for modifying SSA is taken by Hara et al. [24], where an analytically solvable SSA algorithm, called Analytical SSA (AnSSA) is proposed. This is achieved by finding an upper bound to the SSA objective derived by a Taylor approximation close to the optimal value of B under the assumption that $\mathbb{I}_d B\Sigma_e B^\top \mathbb{I}_d^\top = \mathbb{I}_{d\times d}$ for all epoch covariances matrices $\Sigma_e : e \in E$. The optimal projection for the resulting optimization problem,

$$\min_{B:BB^\top=\mathbb{I}} \mathrm{tr}(\tilde{B}(\sum_{e\in E}(\mu_e\mu_e^\top + (\Sigma_e - \bar{\Sigma})\bar{\Sigma}^{-1}(\Sigma_e - \bar{\Sigma})^\top))\tilde{B}^\top) \quad (7)$$

can be found by an eigendecomposition leading to a computationally efficient algorithm. Note that $\tilde{B} = \mathbb{I}_d B$. The AnSSA method corresponds to the Euclidean distance between the covariance matrices if the data is whitened, leading to

$$\min_{B:BB^\top=\mathbb{I}} \mathrm{tr}(\tilde{B}(\sum_{e\in E}(\mu_e\mu_e^\top + (\Sigma_e - \mathbb{I})(\Sigma_e - \mathbb{I})^\top))\tilde{B}^\top). \quad (8)$$

Hara et al. [24] assumed whitening. In the non-whitened case, we consider two possible formulations of the problem, i.e. (7) and

$$\min_{B:BB^\top=\mathbb{I}} \mathrm{tr}(\tilde{B}(\sum_{e\in E}(\mu_e\mu_e^\top + (\Sigma_e - \bar{\Sigma})(\Sigma_e - \bar{\Sigma})^\top))\tilde{B}^\top). \quad (9)$$

We will denote (8) as EuSSA in our evaluation and call the algorithm based on (7) AnSSA. For completeness we

evaluate both options although we consider both to be very similar. Note that the Euclidean distance, while simple, has disadvantages as a metric on the space of covariance matrices, e.g. the swelling effect [45], [46], in which an interpolated matrix has a larger determinant than the two matrices it was interpolated from.

Our approach, introduced in Section III combines both advantages. Namely, we formulate an algorithm which is solvable by an eigenvalue problem while at the same time it adopts a suitable metric reflecting the geometric structure of covariance matrices.

B. Metrics for covariance matrices

Several non-Euclidean distance functions for the space of positive definite covariance matrices have been considered recently, among them for example the aforementioned affine invariant Riemann metric (Horev et al. [7]) and the log-Euclidean metric introduced by Arsigny et al. [46]

$$d_{log}(A,B) = ||\log A - \log B||. \tag{10}$$

Motivated by the log-Euclidean, several different metrics based on simple transformations of the arguments were proposed (see Pigoli et al. [47]). A special class of these are distances based on a decomposition of the form $A = LL^\top$. These include the distances based on the matrix square root

$$d_r(A,B) = ||\sqrt{A} - \sqrt{B}||_{\text{fro}}. \tag{11}$$

Several decompositions of the form $LL^\top = A$ can be defined, since for an arbitrary decomposition L, RL for any rotation R leads to another decomposition. Besides the aforementioned Square root, the Cholesky decomposition is a possible choice, leading to

$$d_c(A,B) = ||L_A - L_B||_{\text{fro}}. \tag{12}$$

for L_A denoting the Cholesky decomposition of the covariance matrix A. Another option is the size-and-shape metric or Procrustes metric [48], which is defined as

$$d_S(A,B) = \inf_{R:RR^\top=\mathbb{I}} ||\sqrt{A} - \sqrt{B}R||_{\text{fro}} \tag{13}$$

and chooses the rotation that optimally aligns both matrices. The geometries induced by the Procrustes metric are called shape and size spaces and were investigated for the topic of landmark analysis by [49].

An important concept for these covariance estimators is that of the Fréchet-mean of a set of covariance matrices, since the geometry induced by the respective distance strongly affects what the mean of a set of covariance matrices is. The sample Fréchet-mean for a given distance d and a set of covariance matrices $\Sigma_e : e \in E$ is given by

$$\bar{\Sigma} = \inf_{\Sigma} \sum_{e=1}^{n} d(\Sigma_e, \Sigma)^2. \tag{14}$$

In some cases, like for the root and log distances, the Fréchet-mean can be derived analytically, e.g. for the root distance we obtain

$$(\bar{\Sigma}_R)^{\frac{1}{2}} = \frac{1}{|E|} \sum_{E} (\Sigma_E)^{\frac{1}{2}}. \tag{15}$$

In the case of the affine invariant Riemann metric, the mean is calculated by a gradient descent algorithm, while the mean due to the Procrustes metric is usually found by the generalized Procrustes algorithm [48]. For further reading on distance for covariance matrices, we refer to [50].

III. Wasserstein SSA

In this section, we propose *Wasserstein SSA* (WaSSA), and its exact and approximate solvers.

A. Wasserstein distance

The Wasserstein or Earth Mover's Distance is a metric defined on probability distributions which arises from the field of optimal transport as the solution to the Monge-Kantorovich-transportation problem (see Villani [51]). The Wasserstein distance indicates the optimal transport plan subject to a cost of transferring "mass" between two points in a metric space. The cost function used in this work will be L2-loss and accordingly the Wasserstein-2 distance, which is defined as

$$W_2(\mu,\nu) = \left(\inf_{\pi} \int_{M\times M} d(x,y)^2 d\pi(x,y) \right)^{\frac{1}{2}} \tag{16}$$

where the infimum is over all couplings π on $M \times M$ whose marginals are the probability distributions μ and ν. If μ, ν are assumed to be Gaussian measures with covariance matrices U,V and zero mean, the optimal transport plan can be derived analytically as

$$T = U^{-\frac{1}{2}} (U^{\frac{1}{2}} V U^{\frac{1}{2}})^{\frac{1}{2}} U^{-\frac{1}{2}}, \tag{17}$$

and with $W(t)$ defined as

$$W(t) = ((1-t)\mathbb{I} + tT)V((1-t)\mathbb{I} + tT), \tag{18}$$

for $t \in [0,1]$, $\mathcal{N}(W(t))$ is a geodesic between both Gaussian measures with respect to the Wasserstein-2 distance [17]. In case of non-zero means m_1, m_2, we can add an interpolation $tm_1 + (1-t)m_2$ to get the geodesic. The (squared) $L2$-Wasserstein distance between these two Gaussian distributions $N(m_1, V)$ and $N(m_2, U)$ is given by

$$\begin{aligned} &D_{\mathrm{W}_2}(\mathcal{N}(m_1,V) \parallel \mathcal{N}(m_2,U)) \\ &= |m_1 - m_2|^2 + \operatorname{tr}(V) + tr(U) - 2\operatorname{tr}((U^{\frac{1}{2}} V U^{\frac{1}{2}})^{\frac{1}{2}}). \end{aligned} \tag{19}$$

This particular case of Wasserstein distance defines a metric on the space of positive semidefinite matrices, and induces a Riemann metric. This distance is also known as the Bures distance in quantum information theory and it corresponds to the Procrustes distance which was defined for size and shape spaces, further it reduces to the Hellinger distance in the diagonal case. While the Kullback-Leibler divergence on Gaussian measures corresponds to a Fisher metric, which is a Riemann metric, the Wasserstein-2 distance induces a different Riemann metric on this space [8], [9].

The works [17], [52] show the equivalence between the L2-Wasserstein distance on Gaussian measures and the Procrustes distance. Álvarez-Esteban et al. [16] recently gave a stable algorithm for the calculation of the Fréchet-mean which is given as Algorithm 1. This algorithm uses pairwise optimal transportation plans to take an simple average in the tangent space. This is repeated until convergence.

Algorithm 1 Wasserstein-Fréchet-Mean solver.

Initialize $\Sigma_0 = \Sigma_R, T = 0, \delta = 1, k = 0$, ϵ some small number
while $\delta > \epsilon$ **do**
  $T = 0$
  for $e \in E$ **do**
    $T = T + \bar{\Sigma}_k^{-\frac{1}{2}}(\bar{\Sigma}_k^{\frac{1}{2}}\Sigma_e\bar{\Sigma}_k^{\frac{1}{2}})^{\frac{1}{2}})\bar{\Sigma}_k^{-\frac{1}{2}}$
  $T = \frac{T}{|E|}$
  $\bar{\Sigma}_{k+1} = T\bar{\Sigma}_k T$
  $\delta = ||\bar{\Sigma}_k - \bar{\Sigma}_{k+1}||_{\text{fro}}$
  $k = k + 1$

B. Algorithms

For this discussion we assume that the data is centered. We propose Wasserstein SSA, which finds a projection to the stationary subspace by solving the optimization problem

$$\min_{B:BB^\top=\mathbb{I}} \sum_{e\in E} D_{\mathrm{W}_2}(\mathcal{N}(\tilde{B}\mu_e, \tilde{B}\Sigma_e\tilde{B}^\top) \parallel \mathcal{N}(0, \tilde{B}\bar{\Sigma}\tilde{B}^\top)). \quad (20)$$

As an advantage of replacing the KL distance in the original SSA problem (3) with the Wasserstein distance, the distance in the sum in the WaSSA problem (20) is analytically given as (19). Therefore, we can solve the WaSSA problem (20) by eigendecomposition, in analogy to AnSSA.

Proposition 1. *The optimal projection for the WaSSA problem* (20) *is given by the leading eigenvectors of*

$$S = \sum_{e=1}^{E} \mu_e\mu_e^\top + \Sigma_e + \bar{\Sigma} - 2(\Sigma_e^{\frac{1}{2}}\bar{\Sigma}\Sigma_e^{\frac{1}{2}})^{\frac{1}{2}}. \quad (21)$$

Proposition 1 follows if we replace the KL-divergence in (3) by the Wasserstein-2 distance to obtain (20) and the arithmetic mean $\bar{\Sigma}$ by the Fréchet-mean defined by (14) and calculated according to Algorithm 1. The resulting objective function is the sum of the Wasserstein-2 distance for epochs which by the linearity of the trace can be combined, leading to the optimization problem

$$\min_{B:BB^\top=\mathbb{I}} \operatorname{tr}\left(\mathbb{I}_d B\left(S\right)B^\top\mathbb{I}_d^\top\right) \quad (22)$$

for S given above. Due to S being positive definite, the optimal rotation B is given by the leading eigenvectors of S corresponding to its smallest eigenvalues. This algorithm does not require an optimization by gradient descent as the original SSA and therefore does not suffer from local minima. AnSSA gave a first analytical version of SSA based by an approximation of the objective, but the resulting algorithm is using the Euclidean distance measure for covariance matrices, which is problematic as we will show experimentally. Algorithm 2 summarizes WaSSA. We empirically observed that Algorithm 1 converges very quickly in 5–7 steps, and the whole procedure is much faster than the previous methods solved by the gradient descent, as shown in more detail in Section IV.

Conveniently, we found that the Wasserstein-Fréchet-mean and the Root-Fréchet-mean are close to each other in our experiments. From this, we propose an approximate variant of WaSSA under an assumption that each covariance matrix Σ_e and the Fréchet-mean $\bar{\Sigma}_W$ is approximately commutative. Under this assumption, the matrix (21) can be approximated as

$$S = \sum_{e=1}^{E} \mu_e\mu_e^\top + (\Sigma_e^{\frac{1}{2}} - \bar{\Sigma}_R^{\frac{1}{2}})(\Sigma_e^{\frac{1}{2}} - \bar{\Sigma}_R^{\frac{1}{2}})^\top. \quad (23)$$

Eq. (23) amounts to measuring the distance between Gaussians by the L2-distance between the means and the matrix-root distance between covariances. Our approximate solver, which we refer to as WaSSA(r) from root, performs the eigendecomposition to the matrix (23), which involves Algorithm 3. We show that WaSSA(r) is even faster than WaSSA by an order of magnitude. Similarity between Wasserstein distance and matrix-root distance implies the robustness of the Wasserstein distance against outliers in the set of covariance matrices. This finding is consistent with our experiments in Section IV, where WaSSA and WaSSA(r) show high robustness against outliers. Our experiments also show that the approximation WaSSA(r) behaves very similar to WaSSA in terms of accuracy, which supports the usefulness of our approximate solver, although the commutativeness assumption might still need further exploration.

Algorithm 2 Wasserstein SSA (WaSSA) solver.

1: Compute by the Wasserstein-Fréchet-mean $\bar{\Sigma}_W$ by Algorithm 1.
2: Calculate eigenvectors u_i, eigenvalues λ_i of S from (21).
3: Take the span of the d eigenvectors u_i correspond to the smallest λ_i as projection to stationary subspace.

Algorithm 3 Wasserstein SSA approximate (WaSSA(r)) solver.

1: Compute Σ_R by (15).
2: Calculate eigenvectors u_i, eigenvalues λ_i of S from (23).
3: Take the span of the d eigenvectors u_i correspond to the smallest λ_i as projection to stationary subspace.

IV. Experimental Evaluation

This section evaluates the proposed WaSSA algorithms on artificial data and on real data from a brain-computer interfacing experiment.

A. Artificial Data Experiment

For the standard SSA algorithm, we relied on the Matlab implementation by Müller et al. [53]. The other SSA variants were implemented by ourselves in Matlab. We use the following abbreviations: Wasserstein-SSA with full Fréchet-mean WaSSA(f) based on (21), the approximation by the root distance WaSSA(r) based on (23), EuclideanSSA (EuSSA) based on (8), and AnalyticalSSA (AnSSA) based on (7). For the Wasserstein algorithm WaSSA(f) we calculated the mean according to the Algorithm 1, where we initialized the

algorithm with the root-mean estimator and the values ϵ = 10e-4. We empirically found Algorithm 1 to typically convergence in 5-7 steps. As an error measure we adopt the subspace error used by [4], which first projects two matrices A, B to the orthogonal manifold and then calculates the error as

$$1 - \text{mean}(\Theta^2) \tag{24}$$

for Θ the singular values of the projections, $A'_O B_O$. All experiments presented were performed for $D = 10$ and an equal number of stationary and non-stationary sources. All experiments in this subsection were performed 250 times, the data was not prewhitened before the application of the SSA algorithms. The mixing matrices in our experiments were chosen randomly from the orthogonal group.

First we analyzed the time-complexity of the different variants (see Fig. 2). While SSA as well as the algorithm for the Fréchet-Mean scale linearly with the number of epochs, the complexity of SSA is significantly higher. For the standard SSA, we ran the gradient descent solver 5 times from random initialization. Fig. 2 clearly shows that WaSSA(f) is faster than SSA by an order of magnitude, and WaSSA(r) further improves by almost the same amount. AnSSA is the fastest. This result shows a clear advantage of the eigenproblem-based SSA variants over the optimization-based SSA algorithm.

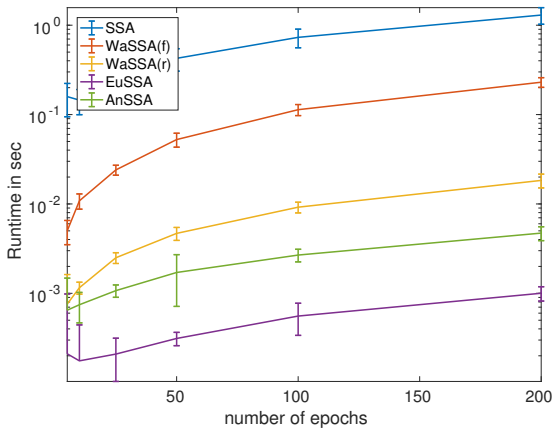

Fig. 2: Computation time comparison.

In a second experiment we evaluated the robustness against outliers. Therefore we created data according to a toy model following Horev et al. [7]. Both sources s_n, s_s are modeled by multivariate Gaussians with zero mean, their covariance have eigenvalues sampled uniformly from $(0, 1)$ which are then randomly rotated. We produce 20 epochs with a total of 10000 datapoints according to this model. To test the robustness towards outliers we add a Gaussian with increasing variance σ^2 to random datapoints with a rate of 0.001 for every datapoint. The results are displayed in Fig. 3. Due to high inter-trial variation we omitted errorbars, instead we tested for significance with an Wilcoxon rank-sum test. We found that the difference from WaSSA(f) and WaSSA(r) to the other baselines are significant for all outlier variances. Additionally, the scatterplot in Fig. 4 displays the results for a particular choice of outlier strength, namely $\sigma^2 = 30$. WaSSA(f) and WaSSA(r) outperform the other algorithms as can be seen. We note that SSA seems to perform well in the ideal cases, while the Euclidean-like algorithms are outperformed for most datapoints.

Overall, this result shows that WaSSA(f) and WaSSA(r) are significantly more robust to outliers than the other SSA variants.

In another experiment we compare the robustness of the different SSA variants to distribution mismatch, i.e., instead of sampling datapoints from a Gaussian distribution[2], we added a Student's t-distribution with 3 degrees of freedom to the datapoints with increasing probability. Note that this simulation is of high practical relevance as the Gaussian assumption will always only approximately hold on real data. Both WaSSA variants outperform the other SSA versions with $p < 0.05$ for intermediate probabilities, but the effect is smaller than for the increasing variance experiment in Fig. 3. We also displayed the scatter plots in Fig. 6.

Another simulation experiment was to test robustness towards time-varying covariance, as was applied in the experiments of Horev et al. [7]. Similar to their experiment we created 50 epochs, but instead of performing this for just a single covariance strength we increase the maximal covariance from 0 to 1. The results are given in Fig. 7. WaSSA turns out to vastly outperform the other methods in this experiment.

In summary, the three robustness experiments show that the Wasserstein based algorithms outperform the other SSA algorithms by a large margin when the data is contaminated by outliers or sampled from a mismatching or (even worse) time-varying distribution. Being robust in these three scenarios is of high importance for practical applications, as also shown in the experiments described in the real data section.

B. Application to change point detection

In this experiment (Fig. 8) we demonstrate that the proposed algorithms provide superior performance in change-point detection experiments. To this we created data following Blythe et al. [22]. Here the non-stationarity is governed by a state model, where each state is defined by a random covariance matrix with eigenvalues sampled randomly from 5 logspaced values between $\frac{1}{p}$ and p for some p which we set to 3. The non-stationary sources are again given by multivariate Gaussians without mean. In this model, after every 100 datapoints, there is a chance $q = 0.1$ for the state to transit into another state. This task is harder to solve than the change point detection in Horev et al. [7] as a number of covariance matrices will look very similar and therefore the information about the distributional changes is harder to extract. In this experiment the ordinary SSA performs well, but the Wasserstein based algorithms come close to its performance. The difference between WaSSA(f/r) and the other analytic algorithms is significant, $p < 0.05$ for all probabilities of change point greater than 0.05.

[2]Note that each of the presented SSA variants models the epoch-wise distribution by a Gaussian.

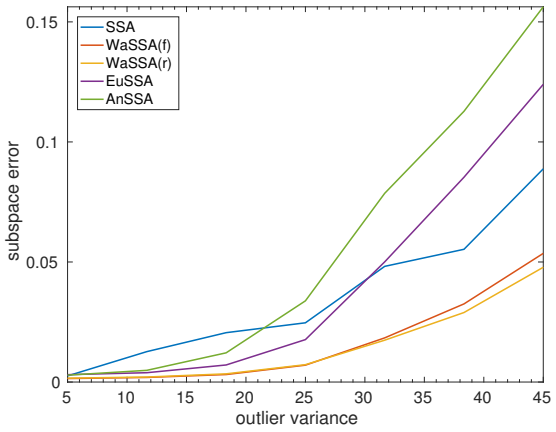

Fig. 3: Robustness against outliers for 0.001 outlier probability.

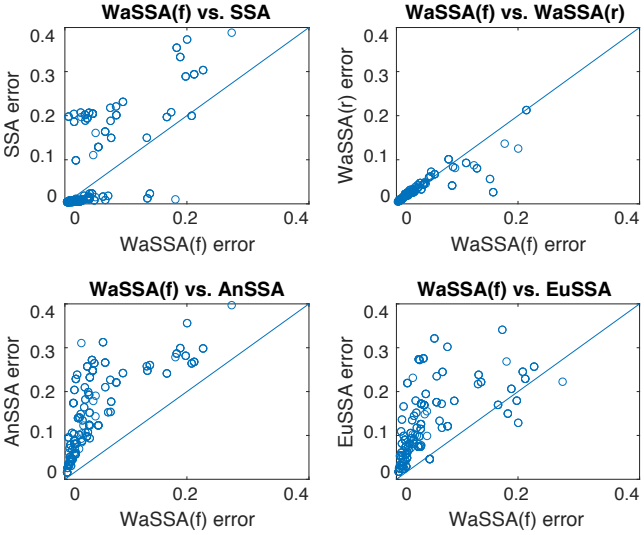

Fig. 4: Scatter plots at outlier variance $\sigma^2 = 30$.

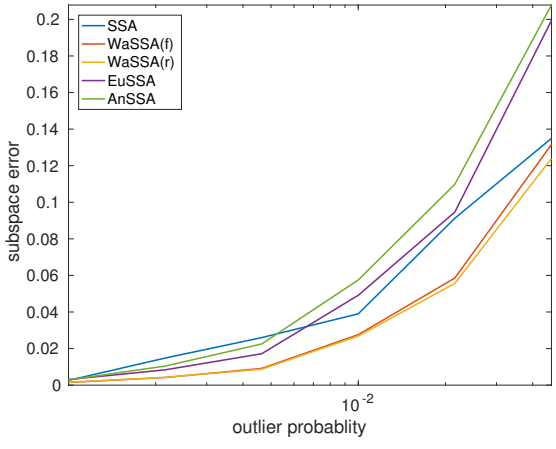

Fig. 5: Robustness against outliers generated from additive Student's t-distribution.

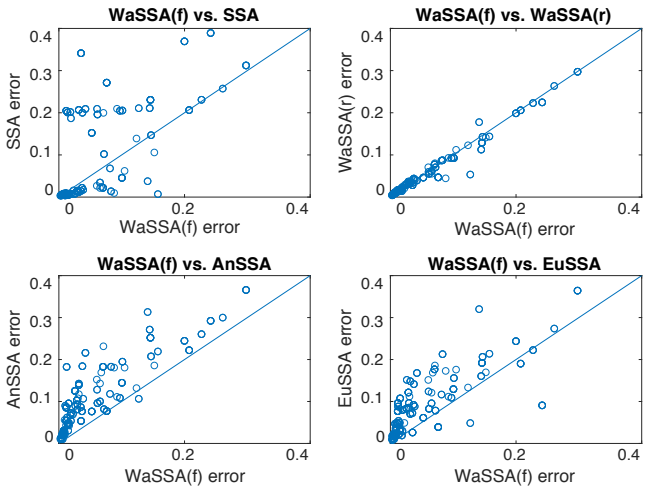

Fig. 6: Scatter plots at outlier probability 0.02.

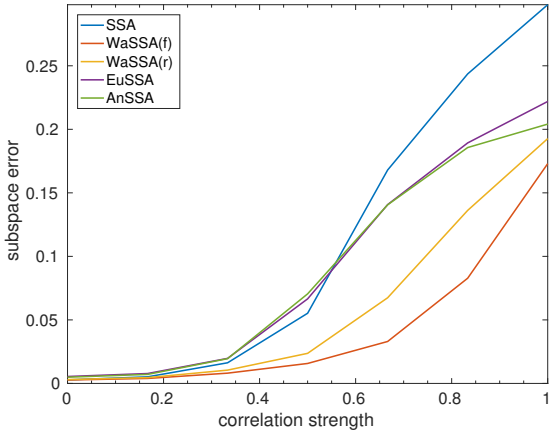

Fig. 7: Robustness under varying correlation.

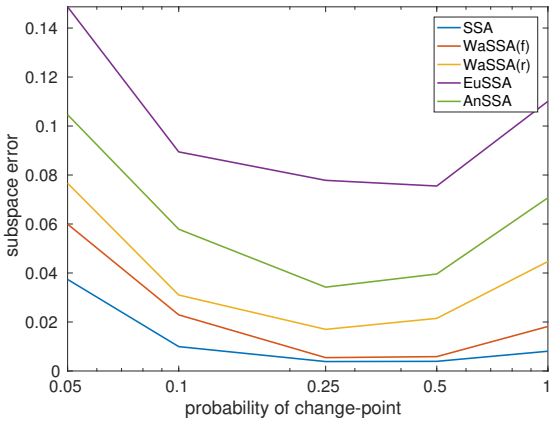

Fig. 8: Accuracy for change point detection experiment.

C. BCI Experiments

This section demonstrates that the proposed SSA method can be used for effectively tackling the non-stationarity problem in spatial filter computation.

1) Dataset and Preprocessing: The evaluation is based on a dataset by Blankertz et al. [54] containing EEG recordings from 80 healthy subjects performing motor imagery tasks with the left and right hand or with the feet. It consists of a calibration session and a feedback session in which subjects had to control a 1D cursor application. Brain activity was recorded from the scalp with multi-channel EEG amplifiers using 119 Ag/AgCl electrodes in an extended 10-20 system sampled at 1000 Hz (downsampled to 100 Hz) with a band-pass from 0.05 to 200 Hz. The two best classes were selected on the calibration data resulting in a training dataset with 150 trials. The subjects performed feedback with these two classes only resulting in a test dataset with 300 trials.

For the offline analysis 62 electrodes densely covering the motor cortex were selected, the data was filtered in 8-30 Hz with a 5-order Butterworth filter and a time segment from 750ms to 3500ms after the trial start was extracted. Six spatial filters computed with different methods were used for feature extraction. A linear discriminant analysis (LDA) classifier was applied to the log-variance features computed on the spatially filtered data. Covariance matrices are normalized by dividing them by their traces and performance is measured as rate of misclassification.

2) Computation of Spatial Filters: In motor imagery based BCIs systems spatial filtering is a crucial step, because it increases the signal-to-noise-ratio and thus simplifies the classification problem. A popular algorithm for computing spatial filters is common spatial patterns (CSP) by Blankertz et al. [33]. It computes spatial filters w by maximizing or minimizing the Rayleigh quotient. One major source of error in the computation of the filters results from the difficulty in proper estimating the class covariance matrices, especially when data is contaminated with artifacts. Furthermore, ignoring the within-class variability and non-stationarity of the signal can result in suboptimal filters.

A variant of CSP which regularizes the filters towards stationary subspaces was proposed in [10]. This stationary CSP (sCSP) algorithm adds a penalty term Δ to the denominator of the Rayleigh quotient, i.e., it maximizes

$$R_1(w) = \frac{w^\top \Sigma_1 w}{w^\top ((1-\lambda)(\Sigma_1+\Sigma_2)+\lambda\Delta) w} \tag{25}$$

$$R_2(w) = \frac{w^\top \Sigma_2 w}{w^\top ((1-\lambda)(\Sigma_1+\Sigma_2)+\lambda\Delta) w} \tag{26}$$

where Σ_1 and Σ_2 are the average covariance matrices of two motor imagery classes and λ controls the strength of the regularization. If Δ is a positive definite matrix, then the resulting optimization problem can be solved very efficiently and has an unique solution. The sCSP algorithm aims to minimize the within-class variability of features measured in terms of absolute differences between the feature in ith trial and the class average. It computes the penalty matrix Δ as

$$\Delta \;=\; \frac{1}{2n}\sum_{c=1}^{2}\sum_{i=1}^{n} \mathcal{F}\left(\Sigma_c^i \;-\; \Sigma_c\right), \tag{27}$$

where $\mathcal{F}$ is an operator to make symmetric matrices positive definite by flipping the sign of all negative eigenvalues. Note that in this formulation sCSP uses a heuristic (namely $\mathcal{F}$) to ensure that Δ is positive definite.

In the following we introduce a very efficient, but also theoretically motivated version of sCSP. In other words, instead of using a heuristic to capture within-class non-stationarity, we will use the root distance as used in the approximation for WaSSA. Since the non-stationarity matrix S computed in (23) is positive definite, it can be directly used as Δ in the sCSP framework in (25) and (26). We call the resulting algorithm *Wasserstein sCSP*.

Analogously, we incorporate the very efficiently computable matrix S of EuSSA based on (8) into the sCSP framework. Also this matrix is positive definite and can be interpreted in terms of Euclidean distances. Therefore, we refer to the resulting algorithm as *Euclidean sCSP*.

3) Results: Figure 9 compares the error rates of CSP, sCSP, Euclidean sCSP and the Wasserstein sCSP method. Each dot represents a subject. Note that the regularization parameter λ has been selected from the set of 10 candidates $\{0, 2^{-8}, ..., 2^{-1}, 2^0\}$ by 5-fold cross-validation on the calibration data (as in [10]). One can clearly see that regularization towards stationary subspaces leads to a decrease in classification error. Almost all subjects benefit from the regularization effect (left scatter plot). The error rate decrease is highly significant with $p < 10^{-4}$ according to the one-sided Wilcoxon signed-rank test. This effect is consistent with what has been reported in previous studies [10], [42], [37].

The Wasserstein distance based measure of non-stationarity also clearly outperforms the Euclidean distance based non-stationarity measure (right scatter plot). Also here the error rate decrease is significant with $p = 0.0018$. One subject clearly benefits from using the Euclidean based regularization (error rate $er = 24.0\%$), but has chance level performance when applying CSP, sCSP or Wasserstein sCSP. Similar improvements can be observed for Wasserstein sCSP. For instance, the error rate of subject 30 decreases up to 20% when computing spatial filters with root distance based regularization, but remains larger than 33% (even for the best parameter) for the other sCSP variants.

Wasserstein sCSP also performs slightly better than the original, heuristic-based sCSP method. On average it leads to a 1.1% lower error rate, however, the error rate decrease is not significant. The improvement over the CSP baseline is highly correlated between Wasserstein sCSP and sCSP with $\rho = 0.66$ (for the best parameters $\rho = 0.88$). Thus, subjects who benefit from sCSP regularization, also benefit from Wasserstein sCSP regularization. For Euclidean sCSP the correlations are much lower, namely $\rho = 0.50$ (for the best parameters $\rho = 0.59$).

In the following we analyze the regularization effects for a particular subject. Figure 10 shows the projected (along largest variance and LDA direction) training (triangles) and

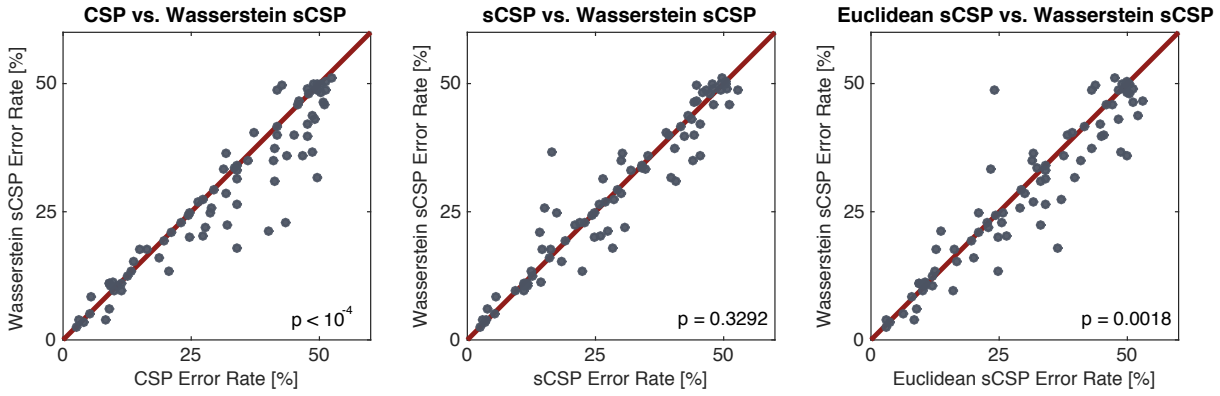

Fig. 9: Scatter plots showing error rates of CSP, sCSP, Euclidean sCSP (x-axis) and the proposed Wasserstein sCSP method (y-axis). Each dot represents a subject and the p-value of the Wilcoxon signed rank test is displayed.

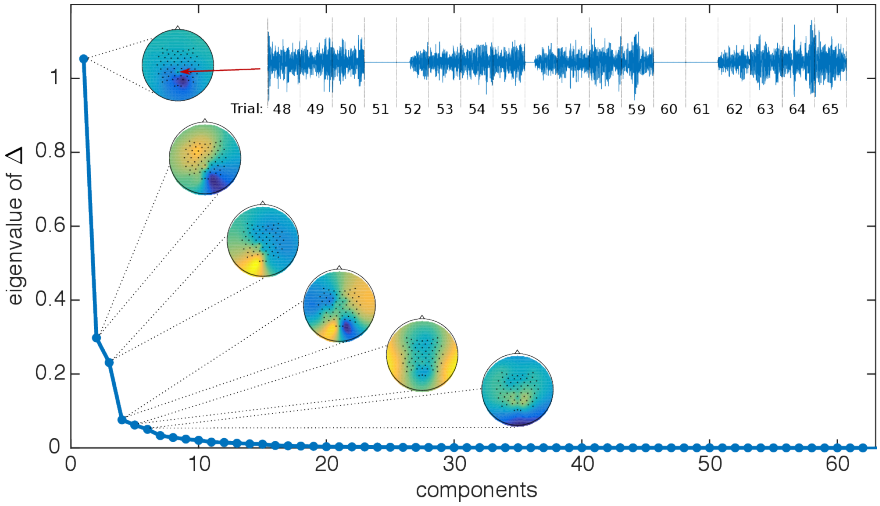

Fig. 11: Eigenvalue spectrum of the penalty matrix Δ of WaSSA. Most non-stationary directions are projected to the scalp. The EEG signal shows that electrode 'CPz' has a problem.

test (circles) features of CSP, sCSP, Euclidean sCSP and Wasserstein sCSP. One can clearly see that CSP features exhibit significant amount of non-stationarity. This results in a large classification error of 34%. Neither the sCSP regularization nor the Euclidean sCSP method solve the non-stationarity problem for this subject. However, the features computed with Wasserstein sCSP show a reduced non-stationarity. With this spatial filter computation method the subject gains BCI control, i.e., the error rate becomes smaller than 30%.

Figure 11 displays the eigenvalue spectrum of the penalty matrix Δ computed with WaSSA. One can see that few eigenvectors capture most of the variation between trials. The top eigenvectors of Δ are displayed on the scalp. The most non-stationary direction shows a clear focus around electrode 'CPz'. The EEG signal next to the scalpplot clearly shows that electrode 'CPz' has a problem in trials 51, 52, 60 and 61. This artifact decreases performance when using CSP, but is regularized out when applying Wasserstein sCSP. The other non-stationarity patterns show activity over parietal areas, which is probably related to visual processing.

V. Discussion

Robustly decomposing non-stationary data is a hard problem. So far subspace decomposition through SSA has been successfully applied to various fields, e.g., geoscience, neuroscience or computer vision. However outliers and noise as well as the need for restarts to find a good SSA solution limit the broader applicability of SSA.

In the present contribution we have therefore extended SSA by providing faster analytical variants and enhanced robustness through usage of Wasserstein distance. The novel Wasserstein SSA algorithm has proven useful on toy and real-world data from the neurosciences. By using Wasserstein, we are able to avoid artifactual effects (e.g. swelling) when interpolating covariance matrices. The scheme proposed has a useful property also beyond SSA for general covariance based subspace estimation algorithms – a subject worthy of further research.

Acknowledgment

We thank for financial support by the German Ministry for Education and Research as Berlin Big Data Center BBDC (funding mark 01IS14013A) and Berlin Center for Machine

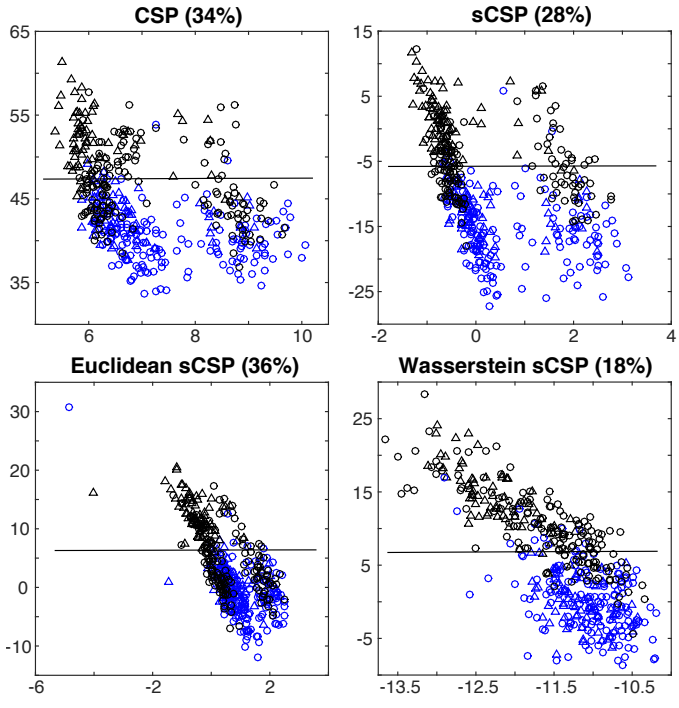

Fig. 10: Training (triangles) and test (circles) features extracted by CSP (upper left), sCSP (upper right), Euclidean sCSP (lower left), Wasserstein sCSP (lower right). The features are projected on largest variance direction (x-axis) and classifier direction (y-axis). Solid line is the decision boundary, color encodes the class label.

Learning BZML (funding mark 01IS18037I). Klaus-Robert Müller acknowledges partial support by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2017-00451, No. 2017-0-01779). Correspondence to Klaus-Robert Müller and Wojciech Samek.

REFERENCES

[1] T. Bouwmans, N. S. Aybat, and E.-H. Zahzah, *Handbook of Robust Low-rank and Sparse Matrix Decomposition: Applications in Image and Video Processing*. CRC Press, 2016.

[2] A. Hyvärinen, J.Karhunen, and E. Oja, *Independant Component Analysis*. John Wiley and Sons, 2002.

[3] G. Blanchard, M. Kawanabe, M. Sugiyama, V. Spokoiny, and K.-R. Müller, "In search of non-gaussian components of a high-dimensional distribution," *Journal of Machine Learning Research*, vol. 7, pp. 247–282, 2006.

[4] P. von Bünau, F. C. Meinecke, F. Király, and K.-R. Müller, "Finding stationary subspaces is multivariate time series," *Physical Review Letters*, vol. 103, no. 21, p. 214101, 2009.

[5] R. Manuca and R. Savit, "Stationarity and nonstationarity in time series analysis," *Physica D: Nonlinear Phenomena*, vol. 99, no. 2, pp. 134–161, 1996.

[6] A. Cichocki, S. Cruces, and S.-i. Amari, "Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences," *Entropy*, vol. 17, no. 5, pp. 2988–3034, 2015.

[7] I. Horev, F. Yger, and M. Sugiyama, "Geometry-aware stationary subspace analysis," *Journal of Machine Learning Research*, vol. 63, pp. 430–444, 2016.

[8] A. Takatsu, "Wasserstein geometry of gaussian measures," *Osaka Journal of Mathematics*, vol. 48, no. 4, pp. 1005–1026, 2011.

[9] R. Bhatia, T. Jain, and Y. Lim, "On the bures–wasserstein distance between positive definite matrices," *Expositiones Mathematicae*, 2018, in press.

[10] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026013, 2012.

[11] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, Eds., *Toward Brain-Computer Interfacing*. MIT Press, 2007.

[12] J. Wolpaw and E. W. Wolpaw, Eds., *Brain-Computer Interfaces: Principles and Practice*. Oxford Univ. Press, 2012.

[13] D. Dowson and B. Landau, "The fréchet distance between multivariate normal distributions," *Journal of Multivariate Analysis*, vol. 12, no. 3, pp. 450–455, 1982.

[14] C. R. Givens and R. M. Shortt, "A class of wasserstein metrics for probability distributions." *Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.

[15] M. Knott and C. S. Smith, "On the optimal mapping of distributions," *Journal of Optimization Theory and Applications*, vol. 43, pp. 39–49, 1984.

[16] P. C. Álvarez Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán, "A fixed-point approach to barycenters in wasserstein space," *Journal of Mathematical Analysis and Applications*, vol. 441, no. 2, pp. 744–762, 2016.

[17] V. Masarotto, V. Panaretos, and Y. Zemel, "Procrustes metrics on covariance operators and optimal transportation of gaussian processes," *arXiv:1801.01990*, 2018.

[18] F. Bachoc, F. Gamboa, J. M. Loubes, and N. Venet, "A gaussian process regression model for distribution inputs," *IEEE Transactions on Information Theory*, pp. 1–1, 2017, in press.

[19] A. Mallasto and A. Feragen, "Learning from uncertain curves: The 2-wasserstein metric for gaussian processes," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5660–5670.

[20] G. Montavon, K.-R. Müller, and M. Cuturi, "Wasserstein training of restricted boltzmann machines," in *Advances In Neural Information Processing Systems 29*, 2016, pp. 3711–3719.

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," *International Conference on Machine Learning (ICML)*, pp. 214–223, 2017.

[22] D. Blythe, P. von Bünau, F. C. Meinecke, and K.-R. Müller, "Feature extraction for change-point detection using stationary subspace analysis," *IEEE Transactions of Neural Networks and Learning Systems*, vol. 23, no. 4, pp. 631–643, 2012.

[23] P. von Bünau, F. C. Meinecke, S. Scholler, and K.-R. Müller, "Finding stationary brain sources in EEG data," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2010, pp. 2810–2813.

[24] S. Hara, Y. Kawahara, T. Washio, P. von Bünau, T. Tokunaga, and K. Yumoto, "Separation of stationary and non-stationary sources with a generalized eigenvalue problem," *Neural Networks*, vol. 33, pp. 7–20, 2012.

[25] M. Baktashmotlagh, M. Harandi, A. Bigdeli, B. Lovell, and M. Salzmann, "Non-linear stationary subspace analysis with application to video classification," *International Conference on Machine Learning (ICML)*, 2013.

[26] D. Panknin, P. von Bünau, M. Kawanabe, F. C. Meinecke, and K.-R. Müller, "Higher order stationary subspace analysis," *Journal of Physics: Conference Series*, vol. 699, no. 1, p. 012021, 2016.

[27] F. J. Király, P. von Bünau, F. Meinecke, D. A. J. Blythe, and K.-R. Müller, "Algebraic geometric comparison of probability distributions," *Journal of Machine Learning Research*, vol. 13, pp. 855–903, 2012.

[28] M. Kawanabe, W. Samek, P. von Bünau, and F. Meinecke, "An information geometrical view of stationary subspace analysis," in *Artificial Neural Networks and Machine Learning - ICANN 2011*, ser. LNCS. Springer, 2011, vol. 6792, pp. 397–404.

[29] F. Castells, P. Laguna, L. Sörnmo, A. Bollmann, and J. M. Roig, "Principal component analysis in ecg signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, p. 074580, 2007.

[30] Y.-P. Lin, P.-K. Jao, and Y.-H. Yang, "Improving cross-day eeg-based emotion classification using robust principal component analysis," *Frontiers in Computational Neuroscience*, vol. 11, p. 64, 2017.

[31] Y. H. Chang, M. Chen, S. Gowda, S. A. Overduin, J. M. Carmena, and C. Tomlin, "Low-rank representation of neural activity and detection of submovements," in *IEEE Conference on Decision and Control (CDC)*, 2013, pp. 2544–2549.

[32] X. Li, H. Zhang, C. Guan, S. H. Ong, K. K. Ang, and Y. Pan, "Discriminative learning of propagation and spatial pattern for motor imagery eeg analysis," *Neural Computation*, vol. 25, no. 10, pp. 2709–2733, 2013.

[33] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Processing Magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[34] F. Lotte and C. Guan, "Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 355–362, 2011.

[35] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013, pp. 1007–1015.

[36] M. Kawanabe, W. Samek, K.-R. Müller, and C. Vidaurre, "Robust common spatial filters with a maxmin approach," *Neural Computation*, vol. 26, no. 2, pp. 349–376, 2014.

[37] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[38] W. Samek, S. Nakajima, M. Kawanabe, and K.-R. Müller, "On robust parameter estimation in brain–computer interfacing," *Journal of Neural Engineering*, vol. 14, no. 6, p. 061001, 2017.

[39] D. B. Thiyam, S. Cruces, J. Olias, and A. Cichocki, "Optimization of alpha-beta log-det divergences and their application in the spatial filtering of two class motor imagery movements," *Entropy*, vol. 19, no. 3, p. 89, 2017.

[40] W. Wojcikiewicz, C. Vidaurre, and M. Kawanabe, "Stationary common spatial patterns: Towards robust classification of non-stationary eeg signals," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 577–580.

[41] ——, "Improving classification performance of bcis by using stationary common spatial patterns and unsupervised bias adaptation," in *Hybrid Artificial Intelligent Systems*, ser. LNCS. Springer, 2011, vol. 6679, pp. 34–41.

[42] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 4, pp. 610–619, 2013.

[43] W. Samek, M. Kawanabe, and C. Vidaurre, "Group-wise stationary subspace analysis - a novel method for studying non-stationarities," in *International Brain-Computer Interface Conference*, 2011, pp. 16–20.

[44] W. Samek, K.-R. Müller, M. Kawanabe, and C. Vidaurre, "Brain-computer interfacing in discriminative and stationary subspaces," *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2873–2876, 2012.

[45] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM Journal on Matrix Analysis and Applications*, vol. 29, no. 1, pp. 328–347, 2007.

[46] ——, "Log-euclidean metrics for fast and simple calculus on diffusion tensors," *Magnetic Resonance in Medicine*, vol. 56, pp. 411–21, 2006.

[47] D. Pigoli, J. Aston, I. Dryden, and P. Secchi, "Distance and inference for covariance operators," *Biometrika*, vol. 101, pp. 409–422, 2014.

[48] J. C. Gower, "Generalized procrustes analysis," *Psychometrica*, vol. 40, pp. 33–51, 1975.

[49] D. G. Kendall, "A survey of the statistical theory of shape," *Statistical Science*, vol. 4, no. 2, pp. 87–99, 1989.

[50] I. Dryden, A. Koloydenko, and D. Zhou, "Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging," *The Annals of Applied Statistics*, vol. 3, pp. 1102–1123, 2009.

[51] C. Villani, "Topics in optimal transportation," *American Mathematical Society*, 2003.

[52] Y. Zemel, "Fréchet means in wasserstein space theory and algorithms," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, 2017.

[53] J. Müller, P. von Bünau, F. Meinecke, F. Király, and K.-R. Müller, "The stationary subspace analysis toolbox," *Journal of Machine Learning Research*, vol. 12, pp. 3065–3069, 2011.

[54] B. Blankertz, C. Sannelli, S. Halder, E. M. Hammer, A. Kübler, K.-R. Müller, G. Curio, and T. Dickhaus, "Neurophysiological predictor of SMR-based BCI performance," *NeuroImage*, vol. 51, no. 4, pp. 1303–1309, 2010.

[55] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis, "Geodesic pca versus log-pca of histograms in the wasserstein space," *SIAM Journal on Scientific Computing*, vol. 40, no. 2, pp. B429–B456, 2018.

Stephan Kaltenstadler received the M.Sc. degrees in physics from Technische Universität München. He is currently with the Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany. His major research interests are in machine learning, information theory and Bayesian methods.

Shinichi Nakajima is a senior researcher in Berlin Big Data Center, Machine Learning Group, Technische Universität Berlin. He received the master degree on physics in 1995 from Kobe university, and worked with Nikon Corporation until September 2014 on statistical analysis, image processing, and machine learning. He received the doctoral degree on computer science in 2006 from Tokyo Institute of Technology. His research interest is in theory and applications of machine learning, in particular, Bayesian learning theory, computer vision, and data mining.

Klaus-Robert Müller (M'12) has been a professor of computer science at Technische Universität Berlin since 2006; at the same time he is co-directing the Berlin Big Data Center. He studied physics in Karlsruhe from 1984 to 1989 and obtained his Ph.D. degree in computer science at Technische Universität Karlsruhe in 1992. After completing a postdoctoral position at GMD FIRST in Berlin, he was a research fellow at the University of Tokyo from 1994 to 1995. In 1995, he founded the Intelligent Data Analysis group at GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. From 1999 to 2006, he was a professor at the University of Potsdam. He was awarded the 1999 Olympus Prize by the German Pattern Recognition Society, DAGM, and, in 2006, he received the SEL Alcatel Communication Award, and, in 2014 he was granted the Science Prize of Berlin awarded by the Governing Mayor of Berlin. In 2012, he was elected to be a member of the German National Academy of Sciences-Leopoldina and in 2017 of the Berlin Brandenburg Academy of sciences. His research interests are intelligent data analysis, machine learning, signal processing, and brain-computer interfaces.

Wojciech Samek (M'13) received a Diploma degree in computer science from Humboldt University of Berlin in 2010 and a Ph.D. degree in machine learning from the Technische Universität Berlin in 2014. In the same year he founded the Machine Learning Group at Fraunhofer Heinrich Hertz Institute which he currently directs. He is associated with the Berlin Big Data Center and was a Scholar of the German National Academic Foundation and a Ph.D. Fellow at the Bernstein Center for Computational Neuroscience Berlin. He was visiting Heriot-Watt University in Edinburgh and the University of Edinburgh from 2007 to 2008. In 2009 he was with the Intelligent Robotics Group at NASA Ames Research Center in Mountain View, CA. His research interests include interpretable machine learning, neural networks, signal processing and computer vision.