Higher order stationary subspace analysis

Danny Panknin[1,†], Paul von Bünau[†], Motoaki Kawanabe[2,‡], Frank C. Meinecke[†], Klaus-Robert Müller[3,†]

[†] Berlin Institute of Technology (TU Berlin), Machine Learning Group, Computer Science
[‡] Advanced Telecommunications Research Institute International (ATR), 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

E-mail: [1] danny.panknin@tu-berlin.de, [2] kawanabe@atr.jp, [3] klaus-robert.mueller@tu-berlin.de
E-mail correspondence to the authors Panknin and Müller.

Abstract. Non-stationarity in data is an ubiquitous problem in signal processing. The recent stationary subspace analysis procedure (SSA) has enabled to decompose such data into a stationary subspace and a non-stationary part respectively. Algorithmically only weak non-stationarities could be tackled by SSA. The present paper takes the *conceptual* step generalizing from the use of first and second moments as in SSA to higher order moments, thus defining the proposed higher order stationary subspace analysis procedure (HOSSA). The paper derives the novel procedure and shows simulations. An obvious trade-off between the necessity of estimating higher moments and the accuracy and robustness with which they can be estimated is observed. In an ideal setting of plenty of data where higher moment information is dominating our novel approach can win against standard SSA. However, with limited data, even though higher moments actually dominate the underlying data, still SSA may arrive on par.

Key words. Stationary subspace analysis, blind source separation, non-stationary data, multivariate time series analysis, dimensionality reduction.

1. Introduction

Non-stationary effects in observed data are a common phenomenon, ranging from the neurosciences (e.g. [1, 2, 3, 4, 5, 6]) to econometrics (e.g. [7]). The causes for temporal distribution changes are manifold. In electroencephalography (EEG) analysis, for instance, non-stationary effects have been attributed to slow drifts of the mental state, the neural response to external stimulation and non-neural (e.g. technical) artefacts (see [8] for discussion).

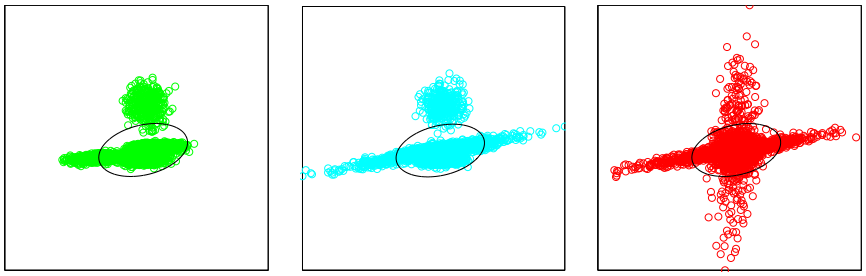
However, detecting distribution changes is difficult in the case where (a) no auxiliary information regarding the distribution changes is available and (b) the observed multivariate data is generated as a mixture of underlying latent factors. In the absence of auxiliary information, such as timing information or a target variable, regression or classification methods are infeasible (unsupervised or explorative setting). Secondly, when each observed variable is generated as a linear combination of latent variables, changes in the joint distribution may not be easily visible in the original coordinates.

Therefore, stationary subspace analysis (SSA) [9] has been proposed to discern the stationary and non-stationary contributions to each observed variable in a completely unsupervised approach. In the SSA model, the observed D-variate time series is generated as a linear mixture

of d_s latent stationary sources $s^{\mathfrak{s}}(t)$ and d_n latent non-stationary sources $s^{\mathfrak{n}}(t)$,

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

where A is an unknown time-constant square mixing matrix of dimension $d_s + d_n$. In the original SSA algorithm [9], this mixture model is inverted by minimizing the distance in distribution of the estimated stationary sources over epochs of the time series (that are defined e.g. by a sliding window). This SSA algorithm has been applied successfully to brain-computer-interfacing (e.g. [9, 10]), myoelectric control [11], computer vision [12], domain adaptation [13], geophysical data analysis [14] and change detection [15, 16]; there is an open-source implementation of the algorithm [17].



(a) Epoch 1 (b) Epoch 2 (c) Epoch 3

Figure 1: The distribution of this bivariate dataset changes between epochs but the mean and the covariance matrix stays the same; the non-stationarity affects only the higher order moments.

However, the SSA algorithm [9] is based on the restricted notion of *weak stationarity*. That is, a time series is considered stationary when its mean and covariance remains constant over time. This means that information in the higher order moments is ignored; e.g. a component with constant mean and covariance but changing kurtosis would be considered stationary. Figure 1 shows an example. In the presence of such a latent factor, the existing SSA algorithm fails to invert the mixture (1). It has been shown that higher moments contain useful information, e.g. in information theory [18, 19], independent component analysis [20] and vision [21].

In this paper, we present the first SSA algorithm which utilizes information from higher order moments based on an approximation of the Kullback-Leibler divergence. For ease of exposition, we restrict ourselves to the third and fourth order moments. Note that our contribution is primarily conceptual in nature. In controlled simulations, we demonstrate the validity of our contribution for higher order information.

The remainder of this paper is organized as follows. In the next Section 2, we briefly review the SSA model and algorithms based on weak stationarity. In Section 3 we derive the proposed method, making use of some concepts of the SSA approach. We compare the performance of SSA and HOSSA in toy examples in Section 4 to reveal drawbacks and advantages of both methods. In the final section, we recapitulate in which situations using HOSSA instead of SSA may improve results and when to stick with SSA.

2. Stationary Subspace Analysis

The aim of SSA is to invert the mixing model (1) given only samples from $x(t)$. That is, we want to find the estimated demixing matrix,

$$\hat{B} = \begin{bmatrix} \hat{B}^{\mathfrak{s}} \\ \hat{B}^{\mathfrak{n}} \end{bmatrix},$$

that separates the contribution of the stationary and the non-stationary sources in the observations. Let $A = \begin{bmatrix} A^{\mathfrak{s}} & A^{\mathfrak{n}} \end{bmatrix}$ be the unknown mixing matrix where the first d_s columns $A^{\mathfrak{s}}$ and the last d_n columns $A^{\mathfrak{n}}$ span the stationary and the non-stationary subspace respectively. In this paper, we assume that the subspace dimensions d_s and d_n are fixed in advance. By using the maximum likelihood version of SSA [10], it may be possible to construct selection procedures of these dimensions from data. However, this is out of the scope of our paper. By applying the demixing matrix $\hat{B}$ to the observations $x(t)$ we obtain estimates $\hat{s}^{\mathfrak{s}}(t)$ and $\hat{s}^{\mathfrak{n}}(t)$ for the underlying latent sources,

$$\begin{bmatrix} \hat{s}^{\mathfrak{s}}(t) \\ \hat{s}^{\mathfrak{n}}(t) \end{bmatrix} = \hat{B}x(t) = \begin{bmatrix} \hat{B}^{\mathfrak{s}} \\ \hat{B}^{\mathfrak{n}} \end{bmatrix} x(t) = \begin{bmatrix} \hat{B}^{\mathfrak{s}}A^{\mathfrak{s}} & \hat{B}^{\mathfrak{s}}A^{\mathfrak{n}} \\ \hat{B}^{\mathfrak{n}}A^{\mathfrak{s}} & \hat{B}^{\mathfrak{n}}A^{\mathfrak{n}} \end{bmatrix} \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}.$$

The inverse of the mixing matrix A^{-1} is a demixing matrix, but it is not unique: any linear transformation *within* the two groups of estimated sources yields another valid separation, because it leaves the stationary resp. non-stationary nature of the sources unchanged. But also the separation into $\mathfrak{s}$- and $\mathfrak{n}$-sources itself is not unique: adding stationary components to a non-stationary source leaves it non-stationary, whereas the converse is not true. That is, the entire set of solutions to the SSA problem is given by the condition $\hat{B}^{\mathfrak{s}}A^{\mathfrak{n}} = 0$. Hence we can identify the true stationary sources derived by applying $\hat{B}^{\mathfrak{s}}$, up to the linear transformation, and the true non-stationary subspace spanned by the columns of $A^{\mathfrak{n}}$, whereas the non-stationary sources and the stationary subspace are not identifiable in general (see also [22]).

The SSA algorithm [9] finds the projections $\hat{B}^{\mathfrak{s}}$ and $\hat{B}^{\mathfrak{n}}$ that minimize resp. maximize the distance in distribution across epochs of the time series. The design of the epoch structure depends on the particular applications; common choices include a sliding window or a chronological segmentation. The distribution in each of the N epochs is approximated by a multivariate Gaussian distribution where $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are the sample mean and sample covariance matrix in the i-th epoch. After a pre-whitening and centering of the average mean and covariance, such that

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_i = 0 \qquad \text{and} \qquad \frac{1}{N}\sum_{i=1}^{N}\hat{\Sigma}_i = I,$$

the distance in distribution across the epochs is measured as the sum of the Kullback-Leibler (KL) divergences D_{KL} between each epoch and the average epoch. KL divergence is the most standard criterion measuring differences of two probability distributions used in statistics, information theory and statistical signal processing. The minimum 0 of this sum is achieved for projections that keep the means and covariances constant. Thus, by minimizing the objective

$$\begin{aligned} \hat{B}^{s} &= \operatorname*{argmin}_{BB^\top = I} \sum_{i=1}^{N} D_{\mathrm{KL}}\Big[\mathcal{N}(B\hat{\mu}_i, B\hat{\Sigma}_i B^\top) \,\Big|\Big|\, \mathcal{N}(0, I)\Big] \\ &= \operatorname*{argmin}_{BB^\top = I} \sum_{i=1}^{N} \left(-\log\det\left(B\hat{\Sigma}_i B^\top\right) + \|B\hat{\mu}_i\|^2\right), \end{aligned}$$

we find a stationary projection due to the weak stationarity assumption. This optimization problem can be solved efficiently by using multiplicative updates with rotation matrices parametrized as matrix exponentials of antisymmetric matrices [9, 23, 17]. Alternatively an algebraic version of SSA has been proposed [22].

3. SSA With Higher Order Moments

Keeping the concepts of second order SSA in mind, we now formulate the task of finding the $\mathfrak{s}$-sources as an optimization problem in an analogous way.

3.1. Establishing an Objective Function

Since we now need to consider differences between the first four moments Gaussian distributions are not adequate, as they are already uniquely determined given the first two moments, such that there exists no degree of freedom to adjust third and fourth moments. Instead we make use of a density approximation, that allows for a direct specification of the first four moments. More specifically, we employ Gram-Charlier expansion which was also applied to ICA for defining non-Gaussianity criteria [24]. Beforehand we will discuss in which form and how to estimate the required moments from data.

3.1.1. Efficient Estimation of Cumulants: Cumulants are a transformed version of simple moments and can therefore be used to represent the corresponding distribution. They are more convenient for our purpose as we can directly find distributions with specified moments setting up a series expansion in terms of the cumulants.

Unlike estimating the first two (multivariate) moments or their variants from data, estimating higher order moments is less commonly known. Given n samples $\{X_1, \ldots, X_n\}$ of a D-dimensional distribution, an unbiased j-th order cumulant estimate $\widehat{\mathcal{K}}^j = [\widehat{\kappa}^{i_1,\ldots,i_j}]_{i_1,\ldots,i_j=1}^{D}$ is in general, component-wise, given by (see e.g. [25], Chapter 4):

$$\widehat{\kappa}^{i_1,\ldots,i_j} = \frac{1}{n} \sum_{k_1,\ldots,k_j=1}^{n} \phi^{k_1\ldots k_j} X_{k_1}^{i_1} \cdot \ldots \cdot X_{k_j}^{i_j}$$

where $\phi^{k_1\ldots k_j} = \frac{(-1)^{\nu-1}}{\binom{n-1}{\nu-1}}$ with $\nu = \#\{k_1, \ldots, k_j\}$, the number of different indices.

Determination of these components in that general form is computationally inefficient, i.e. in $\mathcal{O}(n^j)$. However, rearranging the formulas can reduce the complexity to linear time in n: The j-th order cumulant can be expressed in terms of weighted, already determined lower order cumulants and a linear time sum of j-th order data products. This yields the well known sample mean and covariance for the first and second cumulant. For our purpose we also require the cumulant estimates of third and fourth order, given by

$$\widehat{\kappa}^{r,s,t} = \frac{n}{(n-1)(n-2)} \Big[\sum_{i=1}^{n} X_i^r X_i^s X_i^t - n\widehat{\kappa}^r\widehat{\kappa}^s\widehat{\kappa}^t - (n-1)(\widehat{\kappa}^r\widehat{\kappa}^{s,t} + \widehat{\kappa}^s\widehat{\kappa}^{r,t} + \widehat{\kappa}^t\widehat{\kappa}^{r,s}) \Big]$$

and

$$\begin{aligned}
\widehat{\kappa}^{r,s,t,u} = \frac{1}{(n-1)(n-2)(n-3)} \Big[& n(n+1) \sum_{i=1}^{n} X_i^r X_i^s X_i^t X_i^u - n^2(n+1)\widehat{\kappa}^r\widehat{\kappa}^s\widehat{\kappa}^t\widehat{\kappa}^u \\
& - n(n-1)(n+1)(\widehat{\kappa}^r\widehat{\kappa}^s\widehat{\kappa}^{t,u} + \widehat{\kappa}^r\widehat{\kappa}^t\widehat{\kappa}^{s,u} + \widehat{\kappa}^r\widehat{\kappa}^u\widehat{\kappa}^{s,t} + \widehat{\kappa}^s\widehat{\kappa}^t\widehat{\kappa}^{r,u} + \widehat{\kappa}^s\widehat{\kappa}^u\widehat{\kappa}^{r,t} + \widehat{\kappa}^t\widehat{\kappa}^u\widehat{\kappa}^{r,s}) \\
& - (n-1)^3(\widehat{\kappa}^{r,s}\widehat{\kappa}^{t,u} + \widehat{\kappa}^{r,t}\widehat{\kappa}^{s,u} + \widehat{\kappa}^{r,u}\widehat{\kappa}^{s,t}) \\
& - (n+1)(n-1)(n-2)(\widehat{\kappa}^r\widehat{\kappa}^{s,t,u} + \widehat{\kappa}^s\widehat{\kappa}^{r,t,u} + \widehat{\kappa}^t\widehat{\kappa}^{r,s,u} + \widehat{\kappa}^u\widehat{\kappa}^{r,s,t}) \Big].
\end{aligned}$$

3.1.2. Difference Measurement with KL Divergence: After estimating the first four cumulants for each epoch $\widehat{\kappa}_i^1, \ldots, \widehat{\kappa}_i^4$ and over the whole time $\widehat{\kappa}_0^1, \ldots, \widehat{\kappa}_0^4$ we try, equivalently to second order SSA, to find a transformation of the form $I^{d_s}BW$, which, applied on the time series, gives the d_s estimated stationary components, where I^{d_s} are the upper d_s rows of the unity matrix, W is the already mentioned whitening matrix, and $B \in \mathsf{O}(D)$ is an orthogonal matrix.

Note that cumulants obey the transformation law of contravariant tensors (see e.g. [25], Chapter 2). Therefore it holds that, given a linear transformation $A = [a^{ij}]_{i,j=1}^{d,D} \in \mathbb{R}^{d\times D}$ and an D-dimensional random vector X with j^{th}-order cumulants $\boldsymbol{\kappa}^j = [\kappa^{i_1,\ldots,i_j}]_{i_1,\ldots,i_j=1}^{D}$, the cumulants of the transformed random vector AX, denoted as $\widetilde{\boldsymbol{\kappa}}^j$, are given by

$$\widetilde{\boldsymbol{\kappa}}^j = [\widetilde{\kappa}^{i_1,\ldots,i_j}]_{i_1,\ldots,i_j=1}^{d},$$
$$\widetilde{\kappa}^{i_1,\ldots,i_j} = \sum_{k_1,\ldots,k_j=1}^{D} a^{i_1k_1}\cdot\ldots\cdot a^{i_jk_j}\kappa^{k_1,\ldots,k_j}.$$

Let $\widetilde{\kappa}_i^1, \ldots, \widetilde{\kappa}_i^4$ be the epoch-wise cumulant estimates transformed by $I^{d_s}BW$, and $\widetilde{\kappa}_0^1, \ldots, \widetilde{\kappa}_0^4$ the respective estimates over the whole time.

Note that due to centering, the definition of W and the constraint $BB^\top = I$ it is

$$\widetilde{\kappa}_0^1 = 0 \quad \text{and} \quad \widetilde{\kappa}_0^2 = I.$$

Now, the matrix parameterized by B transforms the time series to the d_s stationary components, when the epoch wise moments are constant over time, i.e. equalling the moments over the whole time. This can be reformulated as the KL divergence being zero between adequately chosen distributions $p_i := p_i(\,\cdot\,; \widetilde{\kappa}_i^1, \widetilde{\kappa}_i^2, \widetilde{\kappa}_i^3, \widetilde{\kappa}_i^4)$, and $p_0 := p_0(\,\cdot\,; 0, I, \widetilde{\kappa}_0^3, \widetilde{\kappa}_0^4)$ respectively, possessing these moments. Formally we want to minimize

$$L(B) = \sum_{i=1}^{N} D_{\mathrm{KL}}\Big[\; p_0 \;\Big|\Big|\; p_i \;\Big]$$

subject to $BB^\top = I$. How such distributions can be constructed is part of the next subsection.

3.1.3. The Gram Charlier Expansion Assume all moments of a distribution exist and were known. When formulating appropriate coefficients $\eta^{i_1\ldots i_j}$, which are functions of the first j moments we can expand the distribution as an infinite series

$$f(\mathbf{t}) = \Phi(\mathbf{t}; \boldsymbol{\kappa}^1, \boldsymbol{\kappa}^2)\left\{1 + \sum_{j=3}^{\infty}\sum_{i_1,\ldots,i_j=1}^{D}\frac{1}{j!}\eta^{i_1\ldots i_j}h^{i_1\ldots i_j}(\mathbf{t})\right\} \tag{2}$$

where $\Phi(\,\cdot\,; \kappa^1, \kappa^2)$ is the *density* function of the Gaussian distribution $\mathcal{N}(\kappa^1, \kappa^2)$ and

$$h^{i_1\ldots i_j}(\mathbf{t}) = \frac{1}{\Phi(\mathbf{t}; \boldsymbol{\kappa}^1, \boldsymbol{\kappa}^2)}\frac{\partial^j\Phi(\mathbf{t}; \boldsymbol{\kappa}^1, \boldsymbol{\kappa}^2)}{\partial x_{i_1}\cdots\partial x_{i_j}}$$

are (multivariate) Hermite polynomials.

Now, that we are interested in the first four moments, we truncate the series (2) to fourth order. Furthermore for simplicity of subsequent estimations, we consider the one-dimensional case, generalizing to arbitrary dimensions later on. The series therefore reduces to

$$f(t) = \Phi(t; \kappa^1, \kappa^2)\left\{1 + \frac{1}{3!}\kappa^3h^3(t) + \frac{1}{4!}\kappa^4h^4(t)\right\} \tag{3}$$

Note, that the corresponding first four moments are still exact, though f is not necessarily a valid probability density anymore.

3.1.4. Approximation of the Objective Function In contrast to SSA the present case exhibits no analytical form of the integral in the KL divergence. Instead we approximate a part of the objective function with a second order Taylor expansion:

$$\log(1+x) \approx x - \frac{x^2}{2}, \tag{4}$$

which then gives:

$$\begin{aligned}
L(B) &= \sum_{i=1}^{N} D_{\mathrm{KL}}\Big[\, p_0 \,\Big|\Big|\, p_i \,\Big] \\
&\overset{(3)}{\simeq} \sum_{i=1}^{N} \int_{\mathbb{R}} \Phi(x;0,1)\left\{1 + \frac{1}{3!}\widetilde{\kappa}_0^3 h_0^3(x) + \frac{1}{4!}\widetilde{\kappa}_0^4 h_0^4(x)\right\} \\
&\quad \times \log\left(\frac{\Phi(x;0,1)\{1 + \frac{1}{3!}\widetilde{\kappa}_0^3 h_0^3(x) + \frac{1}{4!}\widetilde{\kappa}_0^4 h_0^4(x)\}}{\Phi(x;\widetilde{\kappa}_i^1,\widetilde{\kappa}_i^2)\{1 + \frac{1}{3!}\widetilde{\kappa}_i^3 h_i^3(x) + \frac{1}{4!}\widetilde{\kappa}_i^4 h_i^4(x)\}}\right) dx \\
&\overset{(4)}{\simeq} \sum_{i=1}^{N} \int_{\mathbb{R}} \Phi(x;0,1)\left\{1 + \frac{1}{3!}\widetilde{\kappa}_0^3 h_0^3(x) + \frac{1}{4!}\widetilde{\kappa}_0^4 h_0^4(x)\right\} \\
&\quad \times \left\{\log\left(\frac{\Phi(x;0,1)}{\Phi(x;\widetilde{\kappa}_i^1,\widetilde{\kappa}_i^2)}\right) + \frac{1}{3!}\widetilde{\kappa}_0^3 h_0^3(x) + \frac{1}{4!}\widetilde{\kappa}_0^4 h_0^4(x) - \frac{1}{2}\left(\frac{1}{3!}\widetilde{\kappa}_0^3 h_0^3(x) + \frac{1}{4!}\widetilde{\kappa}_0^4 h_0^4(x)\right)^2\right. \\
&\quad \left. - \frac{1}{3!}\widetilde{\kappa}_i^3 h_i^3(x) - \frac{1}{4!}\widetilde{\kappa}_i^4 h_i^4(x) + \frac{1}{2}\left(\frac{1}{3!}\widetilde{\kappa}_i^3 h_i^3(x) + \frac{1}{4!}\widetilde{\kappa}_i^4 h_i^4(x)\right)^2\right\} dx
\end{aligned}$$

Solving integrals of the form above is trivial in the sense that the integrand is a product of a Gaussian density and a polynomial. Unfortunately, the polynomial is large and complicated, such that there is no compact solution which we could present in this paper due to limited space.

3.2. The Minimization

In order to make use of powerful standard analysis gradient descent methods for optimization an adjustment is necessary. Namely standard gradient methods perform additive updates. But the orthogonal matrices form no additive group as they are not closed under addition. So it would be hard to perform such updates ensuring to stay within the set $\mathsf{O}(D)$ of $D \times D$ orthogonal matrices. However, they form a multiplicative group such that multiplicative updates do not suffer from this problem. And since orthogonal matrices can be parameterized with additively closed anti-symmetric matrices we can circumvent the problems making use of the advantages of both sets, performing additive gradient descent in the parameter space of anti-symmetric matrices guaranteeing it to result into a multiplicative update within the orthogonal matrices (see [23] for a detailed version). Note that $\exp(A+B) = \exp(A)\exp(B) = \exp(B)\exp(A)$ is only guaranteed, if A and B commute. Therefore we perform gradient descent steps as a line search along the commutating subgroup generated by $\mathfrak{so}_H = \{tH | t \in \mathbb{R}\}$:

(i) Assume, we are at an initial point $B_0 \in \mathsf{O}(D)$

(ii) Denote by B the desired updated position after a line search step, such that B can be factorized as RB_0 with an orthogonal matrix $R = \exp(tH)$ with positive determinant (i.e. an element in $\mathsf{SO}(D)$) and an anti-symmetric matrix H (i.e. $H^\top = -H$) from the parameter space.

(iii) We can consider the initial position B_0 as a constant and rewrite $L(B) = L_{B_0}(R) = (L_{B_0} \circ \exp)(tH)$. That is, $(L_{B_0} \circ \exp)(0) = L(B_0)$.

One can see that, in order to find H, we simply need to build the gradient of $L_{B_0} \circ \exp$ at position 0 for each step, since we can set $B_0 \leftarrow RB_0$ at the end of each step. The gradient is given by

$$\nabla(L_{B_0} \circ \exp)(0) = (\frac{\partial L_{B_0}}{\partial R}(I))I^\top - I(\frac{\partial L_{B_0}}{\partial R}(I))^\top,$$

where $\frac{\partial L_{B_0}}{\partial R}(I)$ can be derived by differentiating the solution from $L_{B_0}(R)$ along R making use of the fact that $R^\top R = I$. Note that $\nabla(L_{B_0} \circ \exp)(0)$ is antisymmetric. In summary we perform a conjugate gradient method along search directions as proposed by Polak and Ribière [26] as in Algorithm 1, where we denote by **vec** the vectorization of a matrix.

Algorithm 1 Conjugate gradient procedure for HOSSA

1: Choose $B_0 \in \mathsf{O}(D)$ at random and set $k = 0$.
2: **repeat**
3: Determine the next search direction $\mathbf{vec}(H_k) = \frac{g_k^\top (g_k - g_{k-1})}{g_{k-1}^\top g_{k-1}}$ for $k > 0$ and $\mathbf{vec}(H_0) = -g_0$, where $g_k = \mathbf{vec}(\nabla(L_{B_k} \circ \exp)(0))$.
4: Find $t \in \mathbb{R}$ that minimizes $L_{B_k}(\exp(tH_k))$ along $\exp(\mathfrak{so}_{H_k})$.
5: Update $B_{k+1} = \exp(tH_k)B_k$.
6: Set $k \leftarrow k + 1$.
7: **until** convergence

After minimization, the first component of the signal transformed by BW is one of the estimated stationary components. Further components can be extracted by successively applying the algorithm on the remaining estimated components, ending up with a set of orthogonal transformations $b_1, \ldots, b_{d_s}$ (of decreasing dimension). The final transformation is then given by

$$B := B_{d_s} \cdot B_{d_s - 1} \cdot \ldots \cdot B_1,$$

with the embedded transformations

$$B_i := \begin{cases} b_1 & , \quad i = 1 \\ \begin{bmatrix} I_{i-1} & 0 \\ 0 & b_i \end{bmatrix} & , \quad \text{else} \end{cases},$$

where I_k is the k-dimensional identity matrix (further details can be found in [27]).

4. Experiments

The goal of the experiments is to show the validity of the proposed concept in comparison to second order SSA not only in general, but especially in the case, where non-stationarity is not measurable within the first two moments, where we expect second order SSA to fail. In order to analyze the latter case we need a controlled setting fulfilling the specification of the first two moments exhibiting no non-stationarity information. This means, that over all epochs the mean and covariance stay constant. For this purpose we generate a toy dataset with that property.

4.1. Toy Data Generation

Suppose we want to simulate N discriminable epochs of D dimensional data with $d_s < D$ stationary components, where the first two moments stay constant. Let $K = \lceil \log_2 N \rceil$, then we can generate the specified epochs using a Gaussian mixture model with $2K$ components pairing up each two of those. For each pair we generate two states of distribution, that exhibit the property of constant first two moments. Then by choosing one of the two states for all K pairs we can generate up to $2^K > N$ discriminable epochs with the required properties.

Let $\pi_1, \ldots, \pi_{2K} > 0$ be the mixture probabilities the $2K$ components with $\sum_{i=1}^{2K} \pi_i = 1$.

Furthermore let $X_1^0, \ldots, X_{2K}^0$ be independent random vectors with $X_i^0 \sim \mathcal{N}(a_i, \Sigma_i)$ with $a_{K+i} = -\frac{\pi_i}{\pi_{K+i}} a_i, i \in \{1, \ldots, K\}$ and $\Sigma_i = \widetilde{\Sigma}_i + \lambda_i a_i a_i^\top$ with $\widetilde{\Sigma}_i$ positive definite and $\lambda_i = \lambda_{K+i} > 0$.

For the other state of the distributions let $X_1^1, \ldots, X_{2K}^1$ be independent with

$$X_i^1 \sim \mathcal{N}(\sqrt{1+\lambda_i} a_i, \Sigma_i - \lambda_i a_i a_i^\top).$$

Now we choose for the l-th epoch one of the two states for each of the K pairs, i.e. $S^l \in \{0,1\}^K$. Then the Gaussian mixture model of the l-th epoch is given by

$$Y^l = \sum_{i=1}^{K} \left[\mathbb{1}_{\{P=i\}} X_i^{S_i^l} + \mathbb{1}_{\{P=K+i\}} X_{K+i}^{S_i^l} \right],$$

where P indicates the selected component based on the distribution $(\pi_i)_i$, $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function and is independent of all X_i^j. For all $S^l, S^m \in \{0,1\}^K$ the following holds:

$$\mathbb{E}[Y^l] = \mathbb{E}[Y^m] = 0,$$

$$\mathrm{Cov}[Y^l] = \mathrm{Cov}[Y^m] = \sum_{i=1}^{2K} \pi_i \left[\Sigma_i + a_i a_i^\top \right]$$

Now, in order to enforce d_s of the D components to be unchanged in distribution, one can simply set the first d_s entries of each a_i to zero.

4.2. Identifiability

With the toy data from above, we test the subspace identifiability of second order SSA against higher order SSA for different numbers of epochs. As the performance measure we determine the angle between the unique true non-stationary subspace $A^{\mathfrak{n}}$ and its approximation $\hat{A}^n$, given by $\left[\hat{A}^{\mathfrak{s}} \quad \hat{A}^{\mathfrak{n}}\right] = \hat{A} = \hat{B}^{-1}$.

The results in Figure 2 show that, not surprisingly, standard SSA fails to identify the true subspace, invariant of the epoch number. In contrast, higher order SSA converges, in the case of $D = 4$ dimensions, for more than ten epochs to a solution that lies close to the true subspace. Note, that especially the higher order part of the optimization suffers from the limits of approximation such that here the hope of finding and exact solution rather than an approximated one is unfeasible. This can be seen in Figure 2 for the case of $D = 6$ dimensions, where higher order SSA needs much more epochs to converge to a solution, that is close to the true subspace. However, if there is information contained in the first two moments, this information dominates the higher order SSA optimization, letting it act similar to standard SSA.

In fact, we can observe this behavior when we inject information within moments up to second order in a controlled manner. For this we modify the strict model from above, containing no

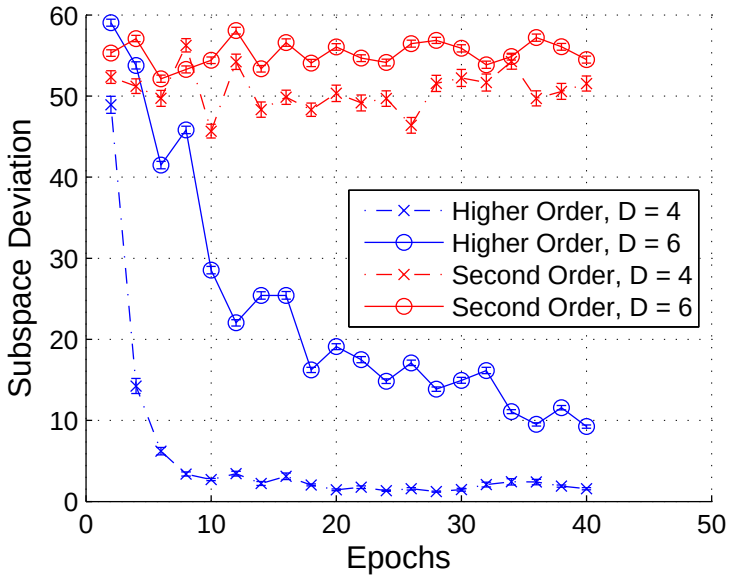


Figure 2: Average deviation of second and higher order SSA from the true solution, measured as the angle between the estimated and the true non-stationary subspace, when the first two moments exhibit no information. Setup: $d_s = 2$, exact (theoretical) Moments, 100 Repetitions.

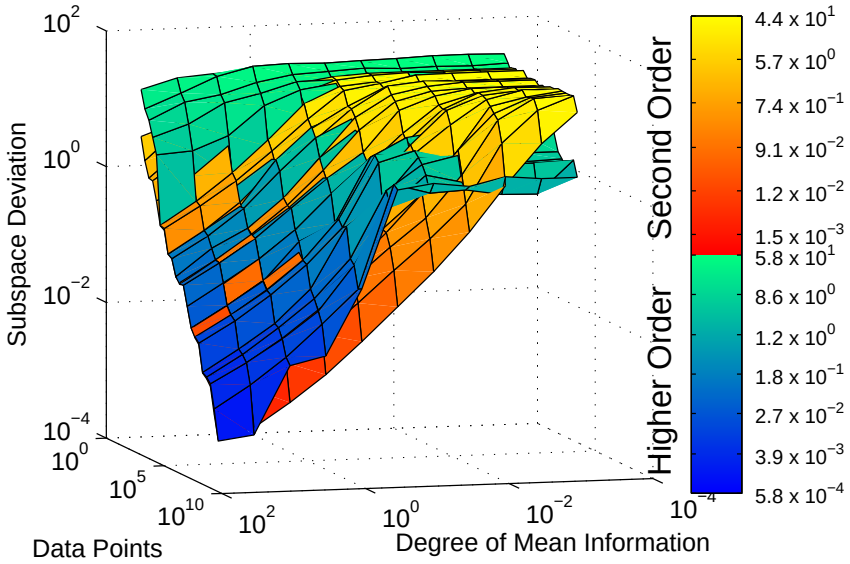


Figure 3: Average deviation of second and higher order SSA from the true solution, measured as the angle between the estimated and the true non-stationary subspace, for a varying amount of data per epoch and information content of first order moments. Setup: $d_s = 2$, $D = 4$, $N = 32$, 20 Repetitions.

information in moments up to second order, by allowing $\lambda_i \neq \lambda_{K+i}$ to a controlled extent. That is, set $\lambda_{K+i} \leftarrow \lambda_i + Cu_i$ for $u_i \sim \mathcal{U}[0,1]$ and a parameter $C \geq 0$, which controls the information content in the first moment. For $C = 0$ we have the strict model with no information in the mean and by increasing C accordingly more information accumulates in the first moment. When fixating a sufficiently large amount of epochs, we can observe in Figure 3 the influence of the parameter C (Degree of Mean Information) for a varying amount of data. Not surprisingly,

second order SSA is more sensitive to the mean-information, starting to benefit for smaller values of C than higher order SSA. For a large C higher order SSA approaches second order SSA despite some approximation related inaccuracy. Since lower order moments are less prone to inexact estimation than higher order moments, for a growing C both methods become more robust with respect to noisy moments or small dataset size. In the area of smaller values of C - where second order SSA is not sensitive enough - we can see that higher order SSA gives usable results in a low noise setting, whereas second order SSA acts completely random. Figure 4 shows slices of Figure 3 for three dataset sizes, namely a too small, a medium and a sufficiently large size, resulting into higher order SSA performing worse, on par or significantly better, respectively, compared to second order SSA.

5. Discussion

Methods to decompose non-stationary data into their stationary and non-stationary subspaces have become important tools for modern data analysis. While the earlier SSA method has made use of the estimation of the second order moments only, the present conceptual extension HOSSA has contributed by employing in addition also higher order moments for decomposition.

Clearly, there is a well-known intrinsic trade-off between robustness and the ability to capture higher order information that we also encounter when comparing SSA and HOSSA. In cases where the information contained in the first two moments is insufficient, a higher order approach can better find the true solution compared to second order SSA – as expected. The proposed method was shown to yield improvement.

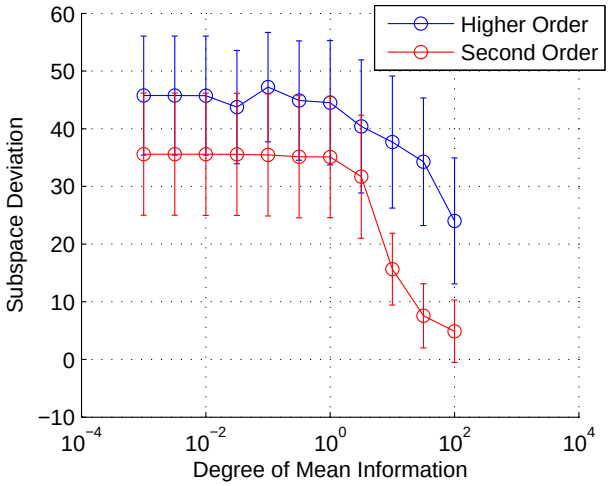
When second and higher order information is both hidden in the signals, we observe that the second order method generally provides a higher robustness. Note, however, that for lower dimensional systems, HOSSA has no significant disadvantage in performance. In higher dimensional examples we have demonstrated that the difficulty in accurately estimating higher order moments becomes the leading reason for a degradation of HOSSA; when the noise level is increased, then this effect becomes amplified. Then, clearly the second order method SSA becomes the more robust estimator for identifying the stationary subspace.

While we might initially think that the additional higher order moments in HOSSA provide additional information allowing an easier identifiability, the experiments show that this potential advantage does not materialize in practice.

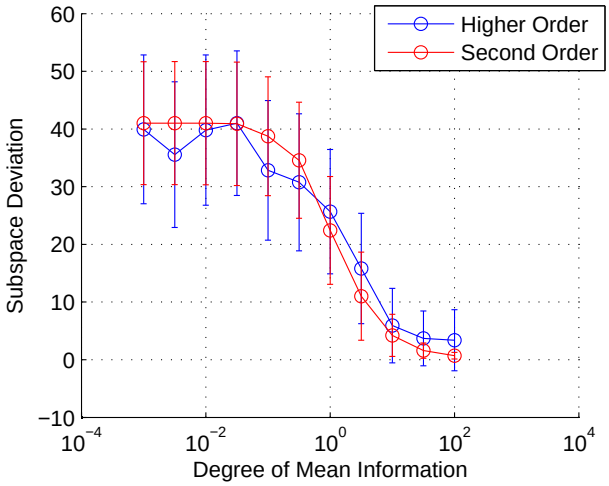
Future work will aim to therefore improve the approximation part of HOSSA, in particular, we will study sampling based approaches. In the independent component analysis field, there have been used other non-Gaussian distributions (e.g. generalized Gaussian distribution in [28]) except for the Gram-Charlier expansion by [24]. More robust implementation of HOSSA might be possible with such a non-Gaussian model.

Acknowledgments

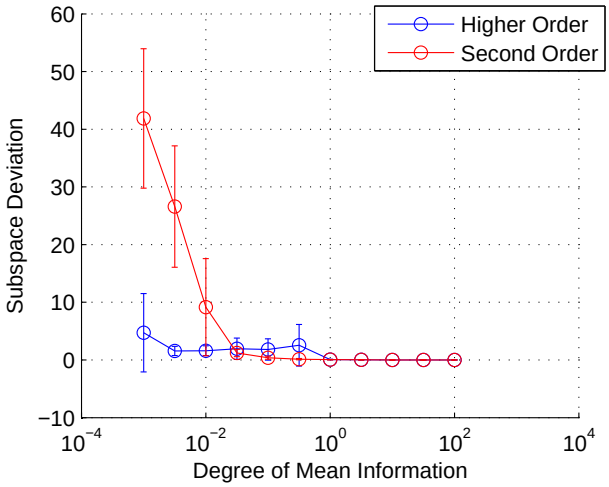
We would like to acknowledge support for this project by the German Ministry of Education and Research (BMBF) through the project ALICE II (Autonomous Learning in Complex Environments II) (01IB15001B) as well as the 'Adaptive BCI' Project, FKZ 01GQ1115, and by the German Research Foundation (DFG) through the project Co-BCI, 'Theoretical concepts for co-adaptive human machine interaction with application to BCI' (MU987/14-1). This work was also supported by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008. MK was supported in part Japan Science and Technology Agency (German-Japanese cooperation program on computational neuroscience), the Ministry of Education, Culture, Sports, Science and Technology (Grant-in-Aid for Scientific Research B, 24300093) and by the Ministry of Internal Affairs and Communications.



(a) Slice for 1000 data points per epoch



(b) Slice for $2.2 \cdot 10^4$ data points per epoch



(c) Slice for $2.2 \cdot 10^9$ data points per epoch

Figure 4: Slices of Figure 3 along fixed data amounts, including approximated 95 % confidence intervals.

References

[1] Shenoy P, Krauledat M, Blankertz B, Rao R P N and Müller K R 2006 Towards adaptive classification for BCI *Journal of Neural Engineering* **3** *1* R13–23
[2] Samek W, Vidaurre C, Müller K R and Kawanabe M 2012 Stationary common spatial patterns for brain-computer interfacing *Journal of Neural Engineering* **9** *2* 026013
[3] Samek W, Meinecke F C and Müller K R 2013 Transferring subspaces between subjects in brain–computer interfacing *IEEE Transactions on Biomedical Engineering* **60** *8* 2289–98
[4] Samek W, Kawanabe M and Müller K R 2014 Divergence-based framework for common spatial patterns algorithms *IEEE Reviews in Biomedical Engineering* **7** 50–72
[5] Vidaurre C, Sannelli C, Müller K R and Blankertz B 2011 Machine-learning-based coadaptive calibration for brain-computer interfaces *Neural Computation* **23** *3* 791–816
[6] Sugiyama M, Krauledat M and Müller K R 2007 Covariate shift adaptation by importance weighted cross validation *Journal of Machine Learning Research* **8** 985–1005
[7] Granger C W J 1981 Some properties of time series data and their use in econometric model specification *Journal of Econometrics* **16** *1* 121-30
[8] Blankertz B, Kawanabe M, Tomioka R, Hohlefeld F, Nikulin V and Müller K R 2008 Advances in neural information processing systems 20 (MIT Press) ed J C Platt et al 113–20
[9] von Bünau P, Meinecke F C, Király F J and Müller K R 2009 Finding stationary subspaces in multivariate time series *Phys. Rev. Lett.* **103** *21* 213101
[10] Kawanabe M, Samek W, von Bünau P and Meinecke F C 2011 Artificial neural networks and machine learning – ICANN 2011 An information geometrical view of stationary subspace analysis *Lecture notes in computer science vol 6792* (Springer Berlin / Heidelberg) 397–404
[11] Hahne J, Dähne S, Hwang H J, Müller K R and Parra L 2015 Concurrent adaptation of human and machine improves simultaneous and proportional myoelectric control *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **23** *4* 618–627
[12] Meinecke F C, von Bünau P, Kawanabe M and Müller K R 2009 Learning invariances with stationary subspace analysis *Lecture notes in computer science vol 6792 IEEE 12th International Conference on Computer Vision Workshops (ICCV Workshops)* 87–92
[13] Hara S, Kawahara Y, Washio T and von Bünau P 2010 Stationary subspace analysis as a generalized eigenvalue problem *Proceedings of the 17th international conference on neural information processing: theory and algorithms* **1** (Springer Berlin / Heidelberg) 422–9
[14] Hara S, Kawahara Y, Washio T, von Bünau P, Tokunaga T and Yumoto K 2012 Separation of stationary and non-stationary sources with a generalized eigenvalue problem *Neural Netw.* **33** (Elsevier Science Ltd.) 7–20
[15] Blythe D A J, von Bünau P, Meinecke F C and Müller K R 2012 Feature extraction for change-point detection using stationary subspace analysis *IEEE Transactions on Neural Networks and Learning Systems* **23** *4* 631–43
[16] Blythe D A J, Meinecke F C, von Bünau P and Müller K R 2013 Explorative data analysis for changes in neural activity *Journal of Neural Engineering* **10** *2* (IOP Publishing) 26018–33
[17] Müller J S, von Bünau P, Meinecke F C, Király F J and Müller K R 2011 The stationary subspace analysis toolbox *Journal of Machine Learning Research* **12** 3065–3069
[18] Jenssen R 2010 Kernel entropy component analysis *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** *5* 847–60
[19] Cover T M and Thomas J A 2006 Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)
[20] Hyvärinen A, Karhunen J and Oja E 2001 Independent component analysis (New York: Wiley)
[21] Hyvärinen A, Hurri J and Hoyer P O 2009 Natural image statistics: a probabilistic approach to early computational vision (Springer publishing company, incorporated)
[22] Király F J, von Bünau P, Meinecke F C, Blythe D A J and Müller K R 2012 Algebraic geometric comparison of probability distributions *Journal of Machine Learning Research* **13** 855–903
[23] Plumbley M D 2005 Geometrical methods for non-negative ICA: manifolds, Lie groups and toral subalgebras *Neurocomputing* **67** (Elsevier Science Publishers B. V.) 161–97
[24] Amari S, Cichocki A and Yang H H 1996 A new learning algorithm for blind signal separation *Advances in Neural Information Processing Systems 8* ed D Touretzky et al (MIT Press) 757–63
[25] McCullagh P 1987 Tensor methods in statistics *Monographs on statistics and applied probability*
[26] Polak E 1971 Computational methods in optimization (Elsevier science & technology books)
[27] von Bünau P 2012 Stationary subspace analysis: towards understanding non-stationary data *Universitätsbibliothek*
[28] Lee T and Lewicki M S 2000 The generalized Gaussian mixture model using ICA *Proceedings of the 2nd International Workshop on Independent Component Analysis* 239–44