An Information Geometrical View of Stationary Subspace Analysis

Motoaki Kawanabe[‡†], Wojciech Samek[⋆†‡],
Paul von Bünau[†], and Frank C. Meinecke[†]

[‡]Fraunhofer Institute FIRST, Kekuléstr. 7, 12489 Berlin, Germany
[†]Berlin Institute of Technology, Franklinsr. 28/29, 10587 Berlin, Germany
motoaki.kawanabe@first.fraunhofer.de, wojwoj@mail.tu-berlin.de
buenau@cs.tu-berlin.de, frank.meinecke@tu-berlin.de

Abstract. Stationary Subspace Analysis (SSA) [3] is an unsupervised learning method that finds subspaces in which data distributions stay invariant over time. It has been shown to be very useful for studying non-stationarities in various applications [5, 10, 4, 9]. In this paper, we present the first SSA algorithm based on a full generative model of the data. This new derivation relates SSA to previous work on finding interesting subspaces from high-dimensional data in a similar way as the three easy routes to independent component analysis [6], and provides an information geometric view.

Keywords: stationary subspace analysis, generative model, maximum likelihood estimation, Kullback-Leibler divergence, information geometry

1 Introduction

Finding subspaces which contain interesting structures in high-dimensional data is an important preprocessing step for efficient information processing. A classical example is Principal Component Analysis (PCA) that extracts low-dimensional subspaces which capture as much variance of the observations as possible. In contrast, more recent approaches consider other criteria than maximum variance. For instance, non-Gaussian component analysis (NGCA) [2, 7], a general semi-parametric framework including projection pursuit [8], extracts *non-Gaussian* structures, e.g. subspaces which reveal clusters or heavy tailed distributions. Another recent direction is colored subspace analysis [13] which searches for linear projections having *temporal correlations*. What all these methods have in common is that i.i.d. (white) Gaussian random variables are considered as noise, and each method seeks to maximize a specific deviation from it to uncover an informative subspace. This fact reminded us of the thought-provoking paper [6] showing three easy routes to Independent Component Analysis (ICA). Independent components can be extracted either by using non-Gaussianity, or by non-whiteness of spectrum, or by non-stationarity [11] of the variance. The

[⋆] née Wojcikiewicz

goal of this paper is to discuss projection methods corresponding to the third route, i.e. based on finding *non-stationary* structures.

Recently, Bünau et al. [3] proposed a novel technique called Stationary Subspace Analysis (SSA) which finds the low-dimensional projections having stationary distributions from high-dimensional observations. This is an instrumental analysis whenever the observed data is conceivably generated as a mixture of underlying stationary and non-stationary components: a single non-stationary source mixed into the observed signals can make the whole data appear non-stationary; and non-stationary sources with low power can remain hidden among strong stationary components. Uncovering these two groups of sources is therefore a key step towards understanding the data. Stationary Subspace Analysis has been applied successfully to biomedical signal processing [5], Computer Vision [10], high-dimensional change point detection [4] and domain adaptation problems [9].

However, the SSA algorithm [3] is based on finding the stationary sources and does not model the non-stationary components, i.e. unlike [11], the SSA algorithm is not derived from any generative model of the observed data. In this paper, we assume a linear mixing model with uncorrelated stationary and non-stationary sources and compute the maximum likelihood estimators of the projections onto both subspaces. That is, we also explicitly model the non-stationary part. The objective function turns out to be a combination of the original SSA objective and a term which penalizes cross-covariances between the stationary and non-stationary sources.

The remainder of this paper is organized as follows. After explaining our model assumptions in Section 2, we derive the objective function of the maximum likelihood approach in Section 3 and provide a geometrical illustration. Then, we present the results of the numerical simulations in Section 4.

2 Block Gaussian Model for SSA

In the SSA model [3], we assume that the system of interest consists of d stationary source signals $\mathbf{s}^{\mathfrak{s}}(t) = [s_1(t), s_2(t), \ldots, s_d(t)]^\top$ (called $\mathfrak{s}$-*sources*) and $D-d$ non-stationary source signals $\mathbf{s}^{\mathfrak{n}}(t) = [s_{d+1}(t), s_{d+2}(t), \ldots, s_D(t)]^\top$ (also $\mathfrak{n}$-*sources*) where the observed signals $x(t)$ are a linear superposition of the sources,

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

and A is an invertible matrix. Note that we do *not* assume that the sources $\mathbf{s}(t)$ are independent. We refer to the spaces spanned by $A^{\mathfrak{s}}$ and $A^{\mathfrak{n}}$ as the $\mathfrak{s}$- and $\mathfrak{n}$-space respectively. The goal is to factorize the observed signals $x(t)$ according to Eq. (1), i.e. to find a linear transformation $\hat{A}^{-1}$ that separates the $\mathfrak{s}$-sources from the $\mathfrak{n}$-sources. Given this model, the $\mathfrak{s}$-sources and the $\mathfrak{n}$-space are uniquely identifiable whereas the $\mathfrak{n}$-sources and the $\mathfrak{s}$-space are not (see [3] for details).

In order to invert the mixing of stationary and non-stationary sources, we divide the time series $x(t)$ of length T into L epochs defined by the index sets

$\mathcal{T}_1, \ldots, \mathcal{T}_L$. Even though the SSA model includes both the stationary and the non-stationary sources, the SSA algorithm [3] optimizes only the stationarity of the estimated stationary sources and does not optimize the non-stationary source estimates: the approach is to find the projection such that the difference between the mean and covariance in each epoch is identical to the average mean and covariance over all epochs.

In this paper, we derive a maximum likelihood estimator for the SSA model under the following additional assumptions.

- The sources $\mathbf{s}$ are independent in time.
- The $\mathfrak{s}$-sources $\mathbf{s}^{\mathfrak{s}}$ are drawn from a d-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}^{\mathfrak{s}}, \Sigma^{\mathfrak{s}})$, which is constant in time
- In each epoch ℓ, the $\mathfrak{n}$-sources $\mathbf{s}^{\mathfrak{n}}$ are drawn from a $(D-d)$-dimensional Gaussian $\mathcal{N}(\boldsymbol{\mu}^{\mathfrak{n}}_\ell, \Sigma^{\mathfrak{n}}_\ell)$, i.e. their distribution varies over epochs
- The $\mathfrak{s}$- and $\mathfrak{n}$-sources are group-wise uncorrelated.

Thus the true distribution of the observed signals $\mathbf{x}(t)$ in epoch ℓ is $\mathcal{N}(\mathbf{m}_\ell, R_\ell)$, where $\mathbf{m}_\ell = A\boldsymbol{\mu}_\ell$, $R_\ell = A\Sigma_\ell A^\top$ and

$$\boldsymbol{\mu}_\ell = \begin{bmatrix} \boldsymbol{\mu}^{\mathfrak{s}} \\ \boldsymbol{\mu}^{\mathfrak{n}}_\ell \end{bmatrix}, \qquad \Sigma_\ell = \begin{bmatrix} \Sigma^{\mathfrak{s}} & 0 \\ 0 & \Sigma^{\mathfrak{n}}_\ell \end{bmatrix}. \tag{2}$$

The unknown parameters of the model are the mixing matrix A, the mean and covariance of the $\mathfrak{s}$-sources $(\boldsymbol{\mu}^{\mathfrak{s}}, \Sigma^{\mathfrak{s}})$ and those of the $\mathfrak{n}$-sources $\{(\boldsymbol{\mu}^{\mathfrak{n}}_\ell, \Sigma^{\mathfrak{n}}_\ell)\}_{\ell=1}^L$.

3 Maximum Likelihood and its Information Geometric Interpretation

In this section, we derive the objective function of the maximum likelihood SSA (MLSSA). Under the block Gaussian model, the negative log likelihood to be minimized becomes

$$\begin{aligned} \mathcal{L}_{\mathrm{ML}} &= -\sum_{\ell=1}^{L} \sum_{t \in \mathcal{T}_\ell} \log p(\mathbf{x}(t) | A, \boldsymbol{\mu}^{\mathfrak{s}}, \boldsymbol{\mu}^{\mathfrak{n}}_\ell, \Sigma^{\mathfrak{s}}, \Sigma^{\mathfrak{n}}_\ell) \\ &= T \log |\det A| + \frac{TD}{2} \log 2\pi + \frac{1}{2} \sum_{\ell=1}^{L} \sum_{t \in \mathcal{T}_\ell} \Big[\log \det \Sigma_\ell \\ &\quad + \mathrm{Tr} \left\{ \Sigma_\ell^{-1} A^{-1} \left(\mathbf{x}(t) - \mathbf{m}_\ell\right) \left(\mathbf{x}(t) - \mathbf{m}_\ell\right)^\top A^{-\top} \right\} \Big]. \end{aligned} \tag{3}$$

It is known that the maximum likelihood estimation can be regarded as the minimum Kullback-Leibler-divergence (KLD) projection [1]. Let

$$\overline{\mathbf{x}}_\ell = \frac{1}{|\mathcal{T}_\ell|} \sum_{t \in \mathcal{T}_\ell} \mathbf{x}(t), \qquad \widehat{R}_\ell = \frac{1}{|\mathcal{T}_\ell|} \sum_{t \in \mathcal{T}_\ell} \left(\mathbf{x}(t) - \overline{\mathbf{x}}_\ell\right) \left(\mathbf{x}(t) - \overline{\mathbf{x}}_\ell\right)^\top \tag{4}$$

be the sample mean and covariance of the ℓ-th chunk, where $|\mathcal{T}_\ell|$ denotes its length. Indeed, the likelihood $\mathcal{L}_{\mathrm{ML}}$ can be expressed as

$$\mathcal{L}_{\mathrm{ML}} = \sum_{\ell=1}^{L} |\mathcal{T}_\ell| \, D_{\mathrm{KL}} \left[\mathcal{N}(\overline{\mathbf{x}}_\ell, \widehat{R}_\ell) \;\middle\|\; \mathcal{N}(\mathbf{m}_\ell, R_\ell) \right] + \text{const.} \tag{5}$$

In the following, we derive the estimators of the moment parameters explicitly as functions of the mixing matrix A which leads to an objective function for estimating A. Since the KL-divergence is invariant under linear transformations, the divergence between the two distributions of the observation $\mathbf{x}$ is equal to that between the corresponding distributions of the sources $\mathbf{s} = A^{-1}\mathbf{x}$, i.e.

$$\begin{aligned} & D_{\mathrm{KL}} \left[\mathcal{N}(\overline{\mathbf{x}}_\ell, \widehat{R}_\ell) \;\middle\|\; \mathcal{N}(\mathbf{m}_\ell, R_\ell) \right] \\ & \qquad = D_{\mathrm{KL}} \left[\mathcal{N}(A^{-1}\overline{\mathbf{x}}_\ell, A^{-1}\widehat{R}_\ell A^{-\top}) \;\middle\|\; \mathcal{N}(\boldsymbol{\mu}_\ell, \Sigma_\ell) \right]. \end{aligned} \tag{6}$$

The central idea of this paper is to regard the KL-divergence as a distance measure in the space of Gaussian probability distributions, which leads to an information geometric viewpoint of the SSA objective. The divergence in Equation (6) is the distance between the empirical distribution of the demixed signals and the true underlying sources. According to our assumption the true model lies on a manifold $\mathcal{M}$ defined by Equation (2), i.e.

$$\mathcal{M} := \left\{ N(\boldsymbol{\mu}, \Sigma) \;\middle|\; \Sigma = \begin{bmatrix} \Sigma^{\mathfrak{s}} & 0 \\ 0 & \Sigma^{\mathfrak{n}} \end{bmatrix} \right\}. \tag{7}$$

Therefore it is convenient to split the divergence (6) into the following two parts,

1. an orthogonal projection D_1 onto the manifold $\mathcal{M}$,
$$D_1 = D_{\mathrm{KL}} \left[\mathcal{N}(A^{-1}\overline{\mathbf{x}}_\ell, A^{-1}\widehat{R}_\ell A^{-\top}) \;\middle\|\; \mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell, \widetilde{\Sigma}_\ell) \right],$$
2. and a component D_2 in the manifold,
$$D_2 = D_{\mathrm{KL}} \left[\mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell, \widetilde{\Sigma}_\ell) \;\middle\|\; \mathcal{N}(\boldsymbol{\mu}_\ell, \Sigma_\ell) \right].$$

This decomposition is illustrated in Figure 1. It is also known as the generalized pythagorean theorem in information geometry [1]. The orthogonal projection onto the manifold $\mathcal{M}$ is given by

$$\widetilde{\boldsymbol{\mu}}_\ell = \begin{bmatrix} \widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{s}} \\ \widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} \left(A^{-1}\right)^{\mathfrak{s}} \overline{\mathbf{x}}_\ell \\ \left(A^{-1}\right)^{\mathfrak{n}} \overline{\mathbf{x}}_\ell \end{bmatrix} = A^{-1}\overline{\mathbf{x}}_\ell, \tag{8}$$

$$\widetilde{\Sigma}_\ell = \begin{bmatrix} \widetilde{\Sigma}_\ell^{\mathfrak{s}} & 0 \\ 0 & \widetilde{\Sigma}_\ell^{\mathfrak{n}} \end{bmatrix} = \begin{bmatrix} \left(A^{-1}\right)^{\mathfrak{s}} \widehat{R}_\ell \left\{ \left(A^{-1}\right)^{\mathfrak{s}} \right\}^{\top} & 0 \\ 0 & \left(A^{-1}\right)^{\mathfrak{n}} \widehat{R}_\ell \left\{ \left(A^{-1}\right)^{\mathfrak{n}} \right\}^{\top} \end{bmatrix}, \tag{9}$$

where $\left(A^{-1}\right)^{\mathfrak{s}}$ and $\left(A^{-1}\right)^{\mathfrak{n}}$ are the projection matrices to the stationary and non-stationary sources respectively. The projection onto the manifold $\mathcal{M}$ does

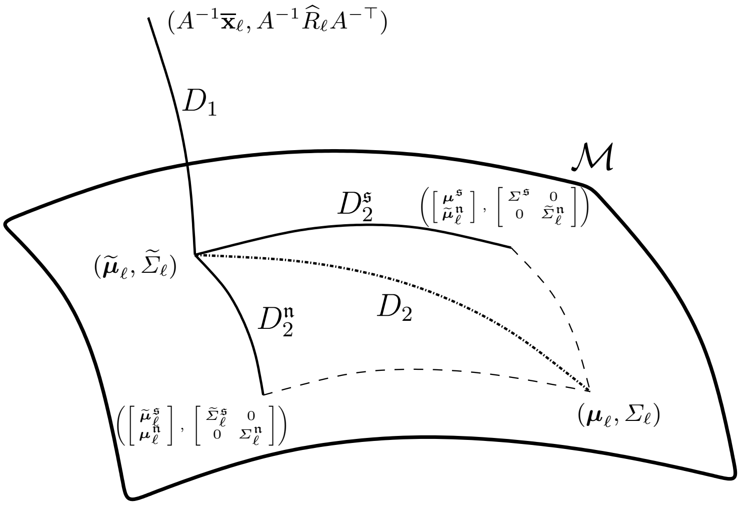
$(A^{-1}\overline{\mathbf{x}}_\ell, A^{-1}\widehat{R}_\ell A^{-\top})$
D_1
$\mathcal{M}$
$D_2^{\mathfrak{s}}$
$\left(\begin{bmatrix}\boldsymbol{\mu}^{\mathfrak{s}}\\ \widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{n}}\end{bmatrix}, \begin{bmatrix}\Sigma^{\mathfrak{s}} & 0\\ 0 & \widetilde{\Sigma}_\ell^{\mathfrak{n}}\end{bmatrix}\right)$
$(\widetilde{\boldsymbol{\mu}}_\ell, \widetilde{\Sigma}_\ell)$
$D_2^{\mathfrak{n}}$
D_2
$\left(\begin{bmatrix}\widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{s}}\\ \boldsymbol{\mu}_\ell^{\mathfrak{n}}\end{bmatrix}, \begin{bmatrix}\widetilde{\Sigma}_\ell^{\mathfrak{s}} & 0\\ 0 & \Sigma_\ell^{\mathfrak{n}}\end{bmatrix}\right)$
$(\boldsymbol{\mu}_\ell, \Sigma_\ell)$

Fig. 1. A geometrical view of the divergence decomposition. The divergence D_1 corresponds to the orthogonal projection onto the manifold of models and D_2 is the divergence on the manifold, which can be further decomposed according to the block diagonal structure of the source covariance matrix.

not affect the mean, it merely sets the off diagonal blocks of the covariance matrix to zero. Conversely, on the manifold $\mathcal{M}$, only the diagonal blocks vary, thus the above projection is orthogonal w.r.t. the Fisher information matrix. For more details, see [1]. Since on $\mathcal{M}$ the stationary and non-stationary sources are independent by definition, we can further decompose the distance D_2 from $\mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell, \widetilde{\Sigma}_\ell)$ to the true distribution $\mathcal{N}(\boldsymbol{\mu}_\ell, \Sigma_\ell)$ into two independent parts,

$$\begin{aligned} D_2 &= D_2^{\mathfrak{s}} + D_2^{\mathfrak{n}} \\ &= D_{\mathrm{KL}}\left[\mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{s}}, \widetilde{\Sigma}_\ell^{\mathfrak{s}}) \;\middle\|\; \mathcal{N}(\boldsymbol{\mu}^{\mathfrak{s}}, \Sigma^{\mathfrak{s}})\right] + D_{\mathrm{KL}}\left[\mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{n}}, \widetilde{\Sigma}_\ell^{\mathfrak{n}}) \;\middle\|\; \mathcal{N}(\boldsymbol{\mu}_\ell^{\mathfrak{n}}, \Sigma_\ell^{\mathfrak{n}})\right]. \end{aligned}$$

This decomposition leads to the following expression for the likelihood,

$$\begin{aligned} \mathcal{L}_{\mathrm{ML}} = \sum_{\ell=1}^{L} |\mathcal{T}_\ell| \Big\{ & D_{\mathrm{KL}}\left[\mathcal{N}(A^{-1}\overline{\mathbf{x}}_\ell, A^{-1}\widehat{R}_\ell A^{-\top}) \;\middle\|\; \mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell, \widetilde{\Sigma}_\ell)\right] \\ & + D_{\mathrm{KL}}\left[\mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{s}}, \widetilde{\Sigma}_\ell^{\mathfrak{s}}) \;\middle\|\; \mathcal{N}(\boldsymbol{\mu}^{\mathfrak{s}}, \Sigma^{\mathfrak{s}})\right] \\ & + D_{\mathrm{KL}}\left[\mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{n}}, \widetilde{\Sigma}_\ell^{\mathfrak{n}}) \;\middle\|\; \mathcal{N}(\boldsymbol{\mu}_\ell^{\mathfrak{n}}, \Sigma_\ell^{\mathfrak{n}})\right] \Big\}. \end{aligned} \tag{10}$$

The moment parameters $(\boldsymbol{\mu}^{\mathfrak{s}}, \Sigma^{\mathfrak{s}})$ of the $\mathfrak{s}$-sources appear only in the second term, while those of the $\mathfrak{n}$-sources $\{(\boldsymbol{\mu}_\ell^{\mathfrak{n}}, \Sigma_\ell^{\mathfrak{n}})\}_{\ell=1}^{L}$ are included only in the third term. Therefore, for fixed A, the moment estimators (denoted with hats) minimizing the likelihood $\mathcal{L}_{\mathrm{ML}}$ can be calculated from $\{(\widetilde{\boldsymbol{\mu}}_\ell, \widetilde{\Sigma}_\ell)\}_{\ell=1}^{L}$:

- for the $\mathfrak{n}$-sources, $(\widehat{\boldsymbol{\mu}}_\ell^{\mathfrak{n}}, \widehat{\Sigma}_\ell^{\mathfrak{n}}) = (\widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{n}}, \widetilde{\Sigma}_\ell^{\mathfrak{n}})$, which means that $D_2^{\mathfrak{n}} = 0$,

- the mean $\widehat{\boldsymbol{\mu}}^{\mathfrak{s}}$ of the $\mathfrak{s}$-sources is the weighted average $\sum_{\ell=1}^{L} \frac{|\mathcal{T}_\ell|}{T} \widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{s}}$ or equivalently $\left(A^{-1}\right)^{\mathfrak{s}} \overline{\mathbf{x}}$, where $\overline{\mathbf{x}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}(t)$ is the global mean,
- the covariance $\Sigma^{\mathfrak{s}}$ of the $\mathfrak{s}$-sources is the projection of the global covariance $\widehat{R} = \frac{1}{T} \sum_{t=1}^{T} \left(\mathbf{x}(t) - \overline{\mathbf{x}}\right) \left(\mathbf{x}(t) - \overline{\mathbf{x}}\right)^\top$ onto the $\mathfrak{s}$-space, i.e. $\left(A^{-1}\right)^{\mathfrak{s}} \widehat{R} \left\{ \left(A^{-1}\right)^{\mathfrak{s}} \right\}^\top$

By substituting these moment estimators, we obtain an objective function to determine the mixing matrix A,

$$\begin{aligned} \mathcal{L}_{\mathrm{ML}} = \sum_{\ell=1}^{L} |\mathcal{T}_\ell| \Big\{ & D_{\mathrm{KL}} \left[\mathcal{N}(A^{-1}\overline{\mathbf{x}}_\ell, A^{-1}\widehat{R}_\ell A^{-\top}) \, \Big\| \, \mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell, \widetilde{\Sigma}_\ell) \right] \\ & + D_{\mathrm{KL}} \left[\mathcal{N}(\widetilde{\boldsymbol{\mu}}_\ell^{\mathfrak{s}}, \widetilde{\Sigma}_\ell^{\mathfrak{s}}) \, \Big\| \, \mathcal{N}(\widehat{\boldsymbol{\mu}}^{\mathfrak{s}}, \widehat{\Sigma}^{\mathfrak{s}}) \right] \Big\} . \end{aligned} \tag{11}$$

Moreover, since the solution is undetermined up to scaling, sign and linear transformations within the $\mathfrak{s}$- and $\mathfrak{n}$-space, we set the estimated $\mathfrak{s}$-sources to zero mean and unit variance by centering and whitening of the data, i.e. we write the estimated demixing matrix as $A^{-1} = B\widehat{W}$ where $\widehat{W} = \widehat{R}^{-1/2}$ is a whitening matrix and B is an orthogonal matrix. Let $\overline{\mathbf{y}}_\ell = \widehat{W}\overline{\mathbf{x}}_\ell$ and $\widehat{V}_\ell = \widehat{W}\widehat{R}_\ell\widehat{W}^\top$ be the mean and covariance of the centerized and whitened data $\mathbf{y}(t) = \widehat{W}\mathbf{x}(t)$ in the ℓ-th epoch. Then, the objective function becomes

$$\begin{aligned} \mathcal{L}_{\mathrm{ML}} = \sum_{\ell=1}^{L} |\mathcal{T}_\ell| \Big\{ & D_{\mathrm{KL}} \left[\mathcal{N}(B\overline{\mathbf{y}}_\ell, B\widehat{V}_\ell B^\top) \, \Big\| \, \mathcal{N}(B\overline{\mathbf{y}}_\ell, \text{b-diag}[B^{\mathfrak{s}}\widehat{V}_\ell \left(B^{\mathfrak{s}}\right)^\top, B^{\mathfrak{n}}\widehat{V}_\ell \left(B^{\mathfrak{n}}\right)^\top]) \right] \\ & + D_{\mathrm{KL}} \left[\mathcal{N}(B^{\mathfrak{s}}\overline{\mathbf{y}}_\ell, B^{\mathfrak{s}}\widehat{V}_\ell \left(B^{\mathfrak{s}}\right)^\top) \, \Big\| \, \mathcal{N}(\mathbf{0}, I_d) \right] \Big\} , \end{aligned} \tag{12}$$

where $B^\top = \left[(B^{\mathfrak{s}})^\top, (B^{\mathfrak{n}})^\top \right]$ and "b-diag" denotes the block diagonal matrix with given sub-matrices. The second term coincides with the original SSA objective. The first term comes from the orthogonality assumption and can be regarded as a joint block diagonalization criterion. The implication of this joint objective is illustrated in Section 4. The objective function (12) can be further simplified as

$$\mathcal{L}_{\mathrm{ML}} = \sum_{\ell=1}^{L} \frac{|\mathcal{T}_\ell|}{2} \left\{ \log\det(\widetilde{\Sigma}_\ell^{\mathfrak{n}}) - \log\det(\overline{\Sigma}_\ell) + (\widetilde{\mu}_\ell^{\mathfrak{s}})^\top \widetilde{\mu}_\ell^{\mathfrak{s}} \right\} + \text{const.} \tag{13}$$

where $\widetilde{\boldsymbol{\mu}}^{\mathfrak{s}} = B^{\mathfrak{s}}\overline{\mathbf{y}}_\ell$, $\widetilde{\Sigma}_\ell^{\mathfrak{s}} = B^{\mathfrak{s}}\widehat{V}_\ell(B^{\mathfrak{s}})^\top$ and $\overline{\Sigma}_\ell := B\widehat{V}_\ell B^\top = A^{-1}\widehat{R}_\ell A^{-\top}$. As in the SSA algorithm [3], we optimize the objective function (12) using a gradient descent algorithm taking into account the Lie group structure of the set of orthogonal matrices [12].

4 Numerical Example

We compare the performance of our novel method with the original SSA algorithm [3]. In particular, we study the case where the epoch covariance matrices are exactly block diagonalizable as this matches the assumptions of our method.

In Figure 2, we compare MLSSA and SSA over varying numbers of epochs using 5 stationary sources and a 20-dimensional input space, i.e. $d = 5$ and $D = 20$. For each epoch we randomly sample a block diagonalizable covariance matrix and a mean vector and mix them with a random, but well-conditioned matrix A. Note that no estimation error is introduced as we sample both moments directly. As performance measure we use the median angle to the true $\mathfrak{n}$-subspace over 100 trials and represent the 25% and 75% quantiles by error bars. For each trial, in order to avoid local minima, we restart the optimization procedure five times and select the solution with the lowest objective function value.

From the results we see that our method yields a much lower error than the original approach, especially when a small number of epochs is used. For instance, when using MLSSA (red line) 10 epochs are sufficient to achieve negligible errors (around 0.01 degree) whereas for the original SSA algorithm (blue line) more than 30 epochs are required. Although it seems that thanks to the first term in Eq. (12), i.e. joint block diagonalization, MLSSA allows a much faster convergence than the original method, more experiments will be needed to obtain a full picture.

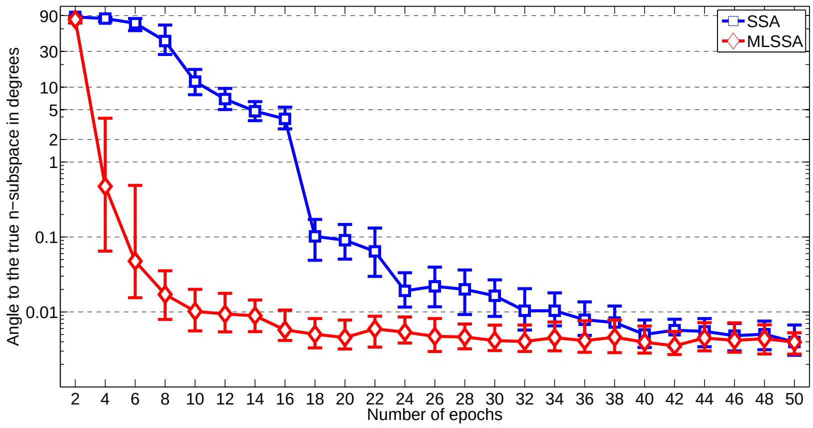

Fig. 2. Performance Comparison between the proposed method (MLSSA) and the original SSA for a varying number of epochs. The red curve shows the performance, measured as the median angle to the true $\mathfrak{n}$-subspace over 100 trials, of our method whereas the blue curve stands for the results obtained with SSA. The 25% and 75% quantiles are represented by the error bars. Note that we use randomly sampled block diagonalizable covariance matrices in our experiments and set the dimensionality of the stationary subspace to 5 (out of 20 dimensions). Our proposed method significantly outperforms SSA in this example, especially when less than 30 epochs are used. MLSSA obviously exploits (by the first term in Eq. (12)) the block diagonalizable structure of the covariance matrices and thus reconstructs the true stationary subspace much faster (i.e. with less epochs) than SSA.

5 Conclusions

In this paper, we developed the first generative version of SSA. The generative model is a block Gaussian model and the objective function of the maximum likelihood approach turns out to be a combination of the original SSA and a joint block diagonalization term. This new derivation not only helps the theoretical understanding of the procedure in the information geometrical framework, but the algorithm also yields competitive results. Moreover, the likelihood formulation allows future extension towards model selection methods and for incorporating prior knowledge.

Acknowledgement The authors acknowledge K.-R. Müller, F. J. Király and D. Blythe for valuable discussions.

References

1. Amari, S.: Differential-geometrical methods in statistics. Lecture notes in statistics, Springer-Verlag, Berlin (1985)
2. Blanchard, G., Sugiyama, M., Kawanabe, M., Spokoiny, V., Müller, K.R.: In search of non-Gaussian components of a high-dimensional distribution. Journal of Machine Learning Research 7, 247–282 (2006)
3. von Bünau, P., Meinecke, F.C., Király, F., Müller, K.R.: Finding stationary subspaces in multivariate time series. Physical Review Letters 103, 214101 (2009)
4. von Bünau, P., Meinecke, F.C., Müller, J.S., Lemm, S., Müller, K.R.: Boosting high-dimensional change point detection with stationary subspace analysis. In: Workshop on Temporal Segmentation at NIPS 2009 (2009)
5. von Bünau, P., Meinecke, F.C., Scholler, S., Müller, K.R.: Finding stationary brain sources in EEG data. In: Proceedings of the 32nd Annual Conference of the IEEE EMBS. pp. 2810–2813 (2010)
6. Cardoso, J.F.: The three easy routes to independent component analysis; contrasts and geometry. In: Proc. ICA 2001. pp. 1–6 (2001)
7. Diederichs, E., Juditsky, A., Spokoiny, V., Schtte, C.: Sparse non-gaussian component analysis. IEEE Trans. Inform. Theory 56, 3033–3047 (2010)
8. Friedman, J.H., Tukey, J.W.: A projection pursuit algorithm for exploratory data analysis. IEEE Trans. Computers 23, 881–890 (1974)
9. Hara, S., Kawahara, Y., Washio, T., von Bünau, P.: Stationary subspace analysis as a generalized eigenvalue problem. In: Proc of 17th international conference on Neural information processing: theory and algorithms - Volume Part I. pp. 422–429. ICONIP'10, Springer-Verlag (2010)
10. Meinecke, F., von Bünau, P., Kawanabe, M., Müller, K.R.: Learning invariances with stationary subspace analysis. In: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. pp. 87 –92 (2009)
11. Pham, D.T., Cardoso, J.F.: Blind separation of instantaneous mixtures of non stationary sources. In: Proc. ICA 2000. pp. 187–192. Helsinki, Finland (2000)
12. Plumbley, M.D.: Geometrical methods for non-negative ica: Manifolds, lie groups and toral subalgebras. Neurocomputing 67(161–197) (2005)
13. Theis, F.: Colored subspace analysis: Dimension reduction based on a signal's autocorrelation structure. IEEE Trans. Circuits & Systems I 57(7), 1463–1474 (2010)