Finding Stationary Subspaces in Multivariate Time Series

Paul von Bünau,[1,*] Frank C. Meinecke,[1,†] Franz C. Király,[2,‡] and Klaus-Robert Müller[1,3,§]

[1]*Machine Learning Group, Computer Science Department, TU Berlin, Germany*
[2]*Institute of Pure Mathematics, University of Ulm, Germany*
[3]*Bernstein Focus Neurotechnology, Berlin, Germany*

(Received 24 March 2009; published 20 November 2009)

Identifying temporally invariant components in complex multivariate time series is key to understanding the underlying dynamical system and predict its future behavior. In this Letter, we propose a novel technique, stationary subspace analysis (SSA), that decomposes a multivariate time series into its stationary and nonstationary part. The method is based on two assumptions: (a) the observed signals are linear superpositions of stationary and nonstationary sources; and (b) the nonstationarity is measurable in the first two moments. We characterize theoretical and practical properties of SSA and study it in simulations and cortical signals measured by electroencephalography. Here, SSA succeeds in finding stationary components that lead to a significantly improved prediction accuracy and meaningful topographic maps which contribute to a better understanding of the underlying nonstationary brain processes.

DOI: 10.1103/PhysRevLett.103.214101 PACS numbers: 05.45.Tp, 02.70.Rr, 07.05.Kf

Discovering and identifying invariances in the dynamics of a physical system is a central task in empirical research. The most fundamental invariances are the laws of physics. In many practical settings, where the aim is to understand a specific dynamical system, one of the most informative invariances is stationarity. Even if the system as a whole does not appear stationary, there may exist subsystems that are stationary. Finding such subsystems is therefore key to characterizing the system—it is here where we contribute with our novel stationary subspace analysis (SSA) technique.

For instance, when measuring cortical activity by electroencephalography (EEG), the corresponding signals usually appear nonstationary. This is partly due to the inherent nonstationary dynamics in the brain. However, even if there exist stationary sources in the brain, these are not discernible in the EEG since we can only measure superpositions of stationary and nonstationary sources. Indeed, we demonstrate that SSA can extract stationary sources from the seemingly nonstationary EEG. This masking of stationary sources by nonstationarities can be observed in a wide range of problems where the observables reflect superpositions of different processes, e.g., in stock market analysis (where volatility in common price drivers conceal long-term relationships [1]) and where only surface measurements are available, e.g., in geophysics. Identifying the stationary subsystem is essential to accurately predict the future behavior of the system, since predictability necessarily relies on the assumption that there are temporal invariances that allow us to transfer knowledge from observed data to the future. Moreover, the decomposition into stationary and nonstationary components contributes to the overall understanding of the system.

The proposed SSA method factorizes a multivariate time series into its stationary and nonstationary components. More precisely, 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 $D - d$ nonstationary 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 [2]. Note that we do *not* assume that the sources $\mathbf{s}(t)$ are independent; see Fig. 1 for an example. 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 [3]. 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

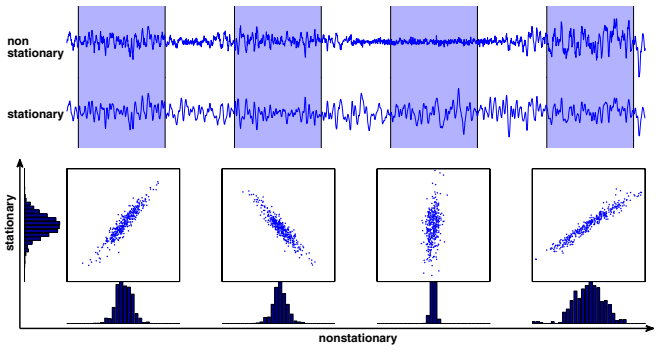

FIG. 1 (color online). Nonstationary and stationary source with time-variable covariance illustrated by amplitude scatter plots for four epochs.

variance by centering and whitening the data; i.e., we write the estimated demixing matrix as $\hat{A}^{-1} = \hat{B}W$ where $W = \mathrm{Cov}(x)^{-1/2}$ is a whitening matrix and $\hat{B}$ is an orthogonal matrix [4].

The rotation part $\hat{B}$ of the demixing is determined using an optimization procedure such that the estimated stationary sources, i.e., the first d components of the estimated sources $\hat{\mathbf{s}}(t) = \hat{B}W\mathbf{x}(t)$, are as stationary as possible. Therefore, we split the data into N consecutive epochs $\mathcal{X}_1, \ldots, \mathcal{X}_N \subset \mathbb{R}^D$ and consider estimated sources as stationary if their joint distribution remains unchanged over all epochs. In order to limit the influence of estimation errors due to finite samples, we characterize the distribution in each epoch $\mathcal{X}_i$ by its empirical mean $\hat{\boldsymbol{\mu}}_i$ and covariance matrix $\hat{\Sigma}_i$ only. To generically compare distributions up to their first two moments, we use the Kullback-Leibler divergence D_{KL} between Gaussians, which is the maximum-entropy distribution that is consistent with the specified moments [7]. Note that we do *not* assume the data to be Gaussian; we only assume that the nonstationarities are visible in the first two moments. Thus, we aim to minimize the KL divergence between the distribution of the estimated $\mathfrak{s}$-sources $\mathcal{N}(\hat{\boldsymbol{\mu}}_i^{\mathfrak{s}}, \hat{\Sigma}_i^{\mathfrak{s}})$ across all epochs, where $\hat{\boldsymbol{\mu}}_i^{\mathfrak{s}} = I^d BW\hat{\boldsymbol{\mu}}_i$, $\hat{\Sigma}_i^{\mathfrak{s}} = I^d BW\hat{\Sigma}_i(I^d BW)^\top$, I^d is the identity matrix truncated to the first d rows and $B^\top B = I$ is the rotation to be determined. The optimization is carried out using multiplicative updates: starting with a random orthogonal matrix, we multiply the current matrix B_0 in each iteration by an orthogonal matrix $B \leftarrow RB_0$. Since we have set the estimated $\mathfrak{s}$-sources to zero mean and unit covariance, minimizing the KL divergence across all epochs is equivalent to minimizing the loss function

$$L_{B_0}(R) = \sum_{i=1}^{N} D_{\mathrm{KL}}[\mathcal{N}(\hat{\boldsymbol{\mu}}_i^{\mathfrak{s}}, \hat{\Sigma}_i^{\mathfrak{s}}) \parallel \mathcal{N}(0, I)]$$
$$= \sum_{i=1}^{N} (-\log\det\hat{\Sigma}_i^{\mathfrak{s}} + \hat{\boldsymbol{\mu}}_i^{\mathfrak{s}\top}\hat{\boldsymbol{\mu}}_i^{\mathfrak{s}}). \quad (2)$$

By parametrizing the orthogonal matrices as the matrix exponentials of antisymmetric matrices, i.e., $R = e^M$, we arrive at a gradient of the shape

$$\frac{\partial L_{B_0}(e^M)}{\partial M}\bigg|_{M=0} = \begin{bmatrix} 0 & Z \\ -Z^\top & 0 \end{bmatrix}. \quad (3)$$

The components of this matrix gradient can be interpreted as infinitesimal rotation angles; i.e., the entry in row i and column j is the angle by which axis i will be rotated towards axis j. Hence, the nonzero part $Z \in \mathbb{R}^{d\times(D-d)}$ corresponds to the rotations between coordinates of the $\mathfrak{s}$- and $\mathfrak{n}$-space [8]. Note that the derivative with respect to the rotations within the two spaces must vanish because they do not change the solution. Thus, we can reduce the number of variables to $d(D-d)$. The optimization is then carried out by conjugate gradient descend [9].

The feasibility of the SSA procedure depends on the number of nonstationary sources $D-d$ and available epochs N. If the number of epochs is too small, we will find spurious stationary directions outside the true $\mathfrak{s}$-subspace due to the limited amount of observed variation in the distributions. For instance, given two distinct Gaussians in two dimensions with covariance matrices Σ_1, Σ_2 and equal means, we can always find two seemingly stationary directions $\mathbf{w} \in \mathbb{R}^2$, i.e., directions where the variances of the projected Gaussians are equal, $\mathbf{w}^\top\Sigma_1\mathbf{w} = \mathbf{w}^\top\Sigma_2\mathbf{w}$ if only the matrix $\Sigma_1 - \Sigma_2$ has an indefinite spectrum. However, the space spanned by these two directions is *not* stationary.

For the general case, it is possible to show that we need

$$N \geq \frac{D-d+1}{2} + 1 \quad (4)$$

epochs to rule out the existence of spurious d-dimensional stationary signals [3]. Because of space limitations, we will only outline the geometrical intuition behind the proof. From N epochs, we obtain $2(N-1)$ equations that restrict the empirical $\mathfrak{s}$-subspace, $N-1$ between the covariance matrices and $N-1$ equations between the epoch means: $\mathbf{w}^\top(\Sigma_n - \Sigma_{n+1})\mathbf{w} = 0$ and $(\boldsymbol{\mu}_n - \boldsymbol{\mu}_{n+1})^\top\mathbf{w} = 0$ with $1 \leq n \leq N-1$. Each of these equations yields a $(D-1)$-dimensional cone or hyperplane. Let $\bar{\mathcal{B}}^{\mathfrak{s}}$ the set-complement of $B^\top$ in D , i.e. all points in data space that do not belong to the true stationary projection. To avoid spurious d-dimensional stationary projections, we have to ensure that the intersection of these hypersurfaces with $\bar{\mathcal{B}}^{\mathfrak{s}}$ has a dimension strictly smaller than d. Since $\bar{\mathcal{B}}^{\mathfrak{s}}$ is of

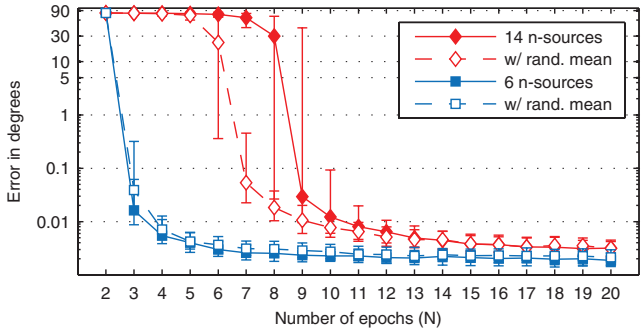

FIG. 2 (color). Median error of SSA measured as the angle between the true and estimated $\mathfrak{n}$-space (vertical axis) for varying numbers of epochs (horizontal axis) on synthetic 20-dimensional data. The red and blue curves show the performance for 14 and 6 nonstationary sources (mixed with a random matrix), respectively. We consider two scenarios: only the covariance matrix of the $\mathfrak{n}$-sources is varied between epochs (solid curves); and both the mean and the covariance matrix are chosen at random (dashed curves). The impact of small sample statistics is excluded by using the exact mean and covariance matrix directly in the SSA procedure. The error bars stretch from the 25% to the 75% quantile estimated over 1000 random realizations of the data.

dimension D, this intersection is at most $[D - 2(N - 1)]$ dimensional and Eq. (4) follows. In the special case of constant epoch means $\boldsymbol{\mu}_i$, Eq. (4) changes to $N > D - d + 1$.

This effect can be observed in the results of the simulations [10] shown in Fig. 2. For 14 nonstationary sources with constant mean (solid red curve), twelve epochs are sufficient to achieve negligible errors, and with six $\mathfrak{n}$-sources (solid blue curve), four epochs are required. Moreover, we see that fewer epochs are required when the mean of the $\mathfrak{n}$-sources varies as well (dashed red line).

Unless the partitioning of the data into epochs is prespecified, we are confronted with a tradeoff: while a certain number of epochs is required to guarantee determinacy, smaller epoch sizes lead to stronger estimation errors. This is illustrated in a second set of simulations with results shown in Fig. 3: for $N \geq 10$, the error is due to small sample statistics. Even though there is no fixed bound to the maximum number of epochs N, the number of required data points per epoch scales with the degrees of freedom in SSA, that is $d(D - d)$. As a rule of thumb, $N \ll \frac{K}{d(D-d)}$ where K is the total number of available samples. In practice, N should be chosen to lie in the plateau between the two bounds (cf. Fig. 3) such that enough variability of the $\mathfrak{n}$-sources is captured *and* the epochs are sufficiently large. Ultimately, the choice of epochs depends on the nature of the analyzed data, e.g., the time scale and strength of the nonstationarities.

Having studied the properties of SSA both theoretically and in simulations, we demonstrate its usefulness in the context of brain computer interfacing (BCI). We have chosen this specific example for two reasons: first, because BCI is a prime example of a difficult task based on multivariate measurements from a nonstationary dynamical system, the active human brain. Second, because in BCI, the effect of SSA is directly visible in improved prediction rates which allows an automated and objective assessment without presupposing any domain-specific expert knowledge.

The goal of BCI is to transmit information directly from the brain to a computer system without the use of peripheral nerves or muscles, e.g., to control a neuroprostheses. The Berlin BCI [11] noninvasively measures cortical activity associated with certain motor imaginations by EEG, in the reported case, imagined movements of the left (i.e., class 1) and right hand (i.e., class 2). In a short calibration phase, the subject is asked to produce motor imagery examples to generate data for both classes [12] which is then used to calibrate a classifier that is able to distinguish both imaginations in the subsequent application phase [11] from the EEG activity. However, there are usually some EEG sources that change from calibration to application. These dynamical changes are one important reason for prediction errors. We indeed find that a restriction to the stationary signals as found by SSA can significantly improve the classification accuracy.

The left panel of Fig. 4 shows the classification error, i.e., the percentage of misclassified EEG epochs in the application phase (i.e., on the test set) when using a state-of-the-art classification method [13] on all 49 EEG channels (red bar) and the error when using only the stationary signals (blue bar). For the presented subject, the SSA preprocessing has reduced the classification error by about one fourth from 16% to 12%. This result is highly significant (Wilcoxon signed rank test on jackknife replicas with p-value $< 10^{-6}$). Note that the SSA decomposition is identified on the calibration set only; i.e., SSA is able to correctly anticipate the structure of possible changes between calibration—and application set without using the latter.

In BCI experiments, one of the most prominent sources of nonstationarity are the so-called alpha oscillations, a

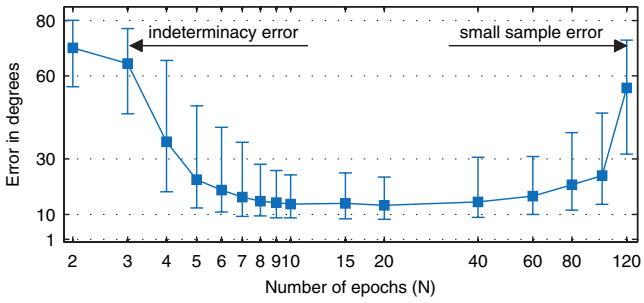

FIG. 3 (color online). Median error of SSA measured as the angle between the true and estimated $\mathfrak{n}$-space (vertical axis) for varying numbers of epochs (horizontal axis) that a fixed number of samples is split into. The data consists of four $\mathfrak{s}$-sources embedded in eight dimensions using a random mixing matrix. The total size of the sample is $K = 1000$; it is composed of seven segments with random length, each with a random covariance matrix for the $\mathfrak{n}$-sources. The error bars stretch from the 25% to the 75% quantile computed over 1000 random realizations of the data.

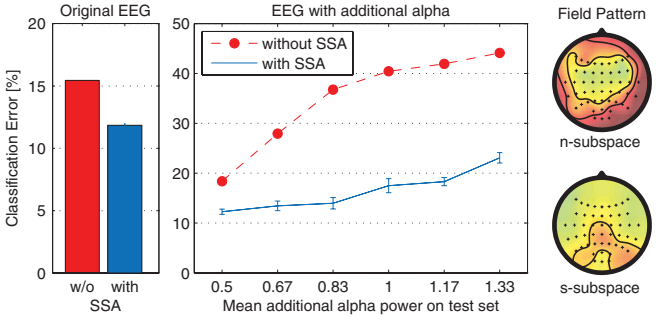

FIG. 4 (color). The left panel shows the classification error on the original EEG BCI data set with and without preprocessing by SSA, the middle panel the evolution of the error for different amounts of alpha oscillations that have been added. In the right panel, the relative differences in signal power between calibration—and test set is plotted as a function of the electrode position. The field pattern shows the head from above, each cross is an electrode. The number of $\mathfrak{s}$-sources is $d = 44$; the number of epochs formed as consecutive trials is $N = 30$. The result is robust with respect to reasonable parameter variations.

strong rhythm in the range between 8 and 12 Hz that is associated with fatigue or tiredness [16]. To quantify the robustness of a classifier in BCI, it is a well established procedure [17] to simulate increasing tiredness over time by superimposing alpha oscillations to the recorded EEG data. This procedure yields a controlled, yet realistic scenario. We extracted typical alpha oscillations from a previous EEG recording by a blind source separation method [5] and added them to the data such that the additional alpha power fluctuated around 0.2 (in units of the average signal strength at electrode ''Oz'') in the calibration set; on the test set, we gradually increased the alpha power. The red curve in the middle panel of Fig. 4 shows deteriorating classification performance for increasing alpha with (blue, solid line) and without (red, dashed line) SSA preprocessing. The figure shows that while the state-of-the-art algorithm quickly reaches error rates close to chance level (i.e., 50%), the SSA preprocessing makes the classification much more robust.

The right panel of Fig. 4 displays the relative differences in signal power between calibration—and application set in both estimated $\mathfrak{n}$ and $\mathfrak{s}$-subspace as field pattern across the scalp. The patterns not only show that the overall signal in the $\mathfrak{s}$-subspace is more stable, but also provide valuable insight into the spatial characteristics of the nonstationarities. In this case, they clearly reflect occipital alpha activity and muscle or ocular artefacts that affect mostly the outer electrodes.

SSA can be applied to a wide range of data because its underlying assumptions are generic, namely, that the observed signals are a linear superposition of sources and that the nonstationarities alter the first two moments. In particular, note that the separation of $\mathfrak{n}$- and $\mathfrak{s}$-sources does not require to impose a restrictive model of the data generating process as in cointegration methods [1]. SSA provides a novel type of analysis which can lead to useful insights into complex physical systems by revealing the location of $\mathfrak{n}$- and $\mathfrak{s}$-sources (if the sensors are spatially distributed as in EEG), identifying meaningful stable relationships between variables (linear combinations in the $\mathfrak{s}$-space correspond to stable equations between variables), and separating $\mathfrak{s}$- from $\mathfrak{n}$-sources for the aim of independent analysis such as prediction or visualization. These are instrumental tasks in many fields of research beyond the neurosciences, e.g., in geophysics (reflection seismology deals with large-scale nonstationary measurements generated by a multitude of sources [18]) and climate research where long-term relationships between variables are difficult to discern due to nonstationary variations in key factors [19]. In any of these domains, SSA may contribute to a better prediction, modeling and thus understanding of the underlying complex physical system.

We gratefully acknowledge support by the Bernstein Cooperation (German Federal Ministry of Education and Science), FKZ 01 GQ 0711; the Bernstein Focus Neurotechnology, and the EU PASCAL2 NoE.

*buenau@cs.tu-berlin.de
†meinecke@cs.tu-berlin.de
‡franz.kiraly@uni-ulm.de
§klaus-robert.mueller@tu-berlin.de

[1] R. F. Engle and C. W. J. Granger, Econometrica **55**, 251 (1987).
[2] It is straightforward to generalize SSA to nonsquare mixing matrices.
[3] See EPAPS Document No. E-PRLTAO-103-014948 for supplementary material. For more information on EPAPS, see http://www.aip.org/pubservs/epaps.html.
[4] This is also standard approach in ICA [5,6].
[5] A. Ziehe, P. Laskov, G. Nolte, and K.-R. Müller, J. Mach. Learn. Res. **5**, 777 (2004).
[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (Wiley, New York, 2001).
[7] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[8] A. Baker, *Matrix Groups* (Springer-Verlag, Berlin, 2002).
[9] M. D. Plumbley, Neurocomputing; Variable Star Bulletin **67**, 161 (2005).
[10] The synthetic data is generated as follows: for each epoch, the mean of the $\mathfrak{n}$-sources is drawn from $\mathcal{N}(0, 2I)$ and the covariance matrix of all sources $\Sigma \in \mathbb{R}^{D\times D}$ is parametrized by Cholesky factors $\Sigma = L^\top L$ with random entries distributed uniformly on $[-0.5, 0.5]$. The entries L_{ij} with $i \leq d$ correspond to the correlation of the $\mathfrak{s}$-sources and are held constant between epochs.
[11] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, NeuroImage **37**, 539 (2007).
[12] In this experiment, the subject was instructed to imagine movements of the right (left) hand for 3 s after seeing the letter *R* (*L*) on a computer screen. The 272 trials of each class were recorded with 49 EEG channels in a random sequence. The data were split chronologically and evenly into a calibration and test set.
[13] The applied state-of-the-art BCI method consists of a dimensionality reduction to 6 dimensions by common spatial patterns [14] followed by linear discriminant analysis [15].
[14] Z. J. Koles, Electroencephalogr Clin Neurophysiol **79**, 440 (1991).
[15] G. Dornhege, J. R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, editors, *Toward Brain-Computer Interfacing* (The MIT Press, Cambridge, MA, 2007).
[16] H. Berger, Archiv fur Psychiatrie und Nervenkrankheiten **87**, 527 (1929).
[17] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, and K.-R. Müller, in *Advances in Neural Information Processing Systems 20*, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (MIT Press, Cambridge, MA, 2008), pp. 113–120.
[18] O. Yilmaz, *Seismic Data Analysis* (Society Of Exploration Geophysicists, Tulsa, OK, 2001).
[19] R. K. Kaufmann and D. I. Stern, J. Geophys. Res. **107**, 4012 (2002).