Stationary Subspace Analysis

Paul von Bünau Frank C. Meinecke Franz J. Király
Klaus-R. Müller

Lecture Machine Learning II

Motivation

- Non-stationarities are ubiquitous in real-world data.
- For example: different training and test distributions
  $\longrightarrow$ hard to generalise.

Motivation

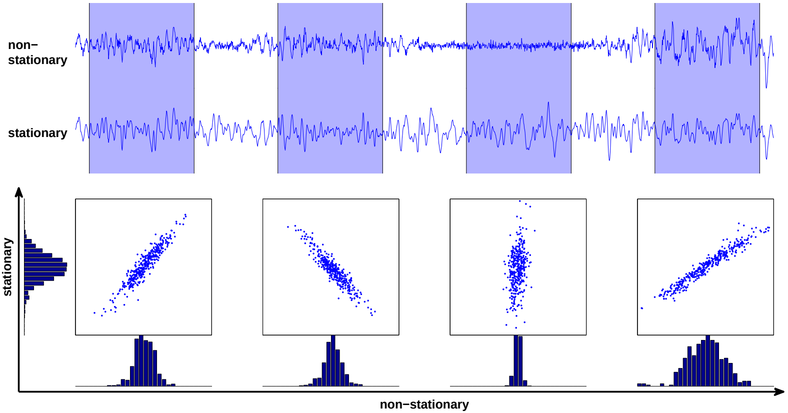
- Non-stationarities are ubiquitous in real-world data.
- For example: different training and test distributions
  $\longrightarrow$ hard to generalise.

Observation

Data generating systems are often only partly non-stationary.

- Removing the non-stationary part can help prediction methods.
- Understanding non-stationary data by factorizing into stationary and non-stationary components.

Stationary and Non-stationary Sources



Generative Model

Assumption

The non-stationarity is confined to a linear subspace of the D-dimensional data space.

Generative Model

Assumption

The non-stationarity is confined to a linear subspace of the D-dimensional data space.

- d stationary source signals $s^{\mathfrak{s}}(t) \in \mathbb{R}^d$

Generative Model

Assumption

The non-stationarity is confined to a linear subspace of the D-dimensional data space.

- d stationary source signals $s^{\mathfrak{s}}(t) \in \mathbb{R}^{d}$
- $D - d$ non-stationary source signals $s^{\mathfrak{n}}(t) \in \mathbb{R}^{(D-d)}$

Generative Model

Assumption

The non-stationarity is confined to a linear subspace of the D-dimensional data space.

- d stationary source signals $s^{\mathfrak{s}}(t) \in \mathbb{R}^d$
- $D - d$ non-stationary source signals $s^{\mathfrak{n}}(t) \in \mathbb{R}^{(D-d)}$
- Observed signals: instantaneous linear superpositions of sources

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

$A^{\mathfrak{s}}$ and $A^{\mathfrak{n}}$ span the stationary and non-stationary subspace respectively.

Goal of Stationary Subspace Analysis

Model

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

Goal of SSA

Given only $x(t)$, find an estimate for the demixing matrix $\hat{B} = \hat{A}^{-1}$ that separates $\mathfrak{s}$-sources from $\mathfrak{n}$-sources.

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

Goal of Stationary Subspace Analysis

Model

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

Goal of SSA

Given only $x(t)$, find an estimate for the demixing matrix $\hat{B} = \hat{A}^{-1}$ that separates $\mathfrak{s}$-sources from $\mathfrak{n}$-sources.

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

Clearly, $\hat{A} = A$ is a solution. But are there other solutions? What are the invariances inherent in the task of separating stationary from non-stationary sources?

Symmetries and Invariances (I)

Let's express the true $A^{\mathfrak{s}}$ and $A^{\mathfrak{n}}$ as linear combinations of the respective estimated subspaces

$$\begin{aligned} A^{\mathfrak{s}} &= \hat{A}^{\mathfrak{s}} M_1 + \hat{A}^{\mathfrak{n}} M_2 \\ A^{\mathfrak{n}} &= \hat{A}^{\mathfrak{s}} M_3 + \hat{A}^{\mathfrak{n}} M_4 \end{aligned}$$

Symmetries and Invariances (I)

Let's express the true $A^{\mathfrak{s}}$ and $A^{\mathfrak{n}}$ as linear combinations of the respective estimated subspaces

$$\begin{aligned} A^{\mathfrak{s}} &= \hat{A}^{\mathfrak{s}} M_1 + \hat{A}^{\mathfrak{n}} M_2 \\ A^{\mathfrak{n}} &= \hat{A}^{\mathfrak{s}} M_3 + \hat{A}^{\mathfrak{n}} M_4 \end{aligned}$$

The composite transformation (true mixing followed by the estimated demixing) reads

$$\begin{bmatrix} \hat{s}^{\mathfrak{s}}(t) \\ \hat{s}^{\mathfrak{n}}(t) \end{bmatrix} = \hat{B} A s(t) = \begin{bmatrix} \hat{B}^{\mathfrak{s}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{s}} A^{\mathfrak{n}} \\ \hat{B}^{\mathfrak{n}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} s(t) = \begin{bmatrix} M_1 & M_3 \\ M_2 & M_4 \end{bmatrix} \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}$$

Symmetries and Invariances (I)

Let's express the true $A^{\mathfrak{s}}$ and $A^{\mathfrak{n}}$ as linear combinations of the respective estimated subspaces

$$\begin{aligned} A^{\mathfrak{s}} &= \hat{A}^{\mathfrak{s}} M_1 + \hat{A}^{\mathfrak{n}} M_2 \\ A^{\mathfrak{n}} &= \hat{A}^{\mathfrak{s}} M_3 + \hat{A}^{\mathfrak{n}} M_4 \end{aligned}$$

The composite transformation (true mixing followed by the estimated demixing) reads

$$\begin{bmatrix} \hat{s}^{\mathfrak{s}}(t) \\ \hat{s}^{\mathfrak{n}}(t) \end{bmatrix} = \hat{B} A s(t) = \begin{bmatrix} \hat{B}^{\mathfrak{s}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{s}} A^{\mathfrak{n}} \\ \hat{B}^{\mathfrak{n}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} s(t) = \begin{bmatrix} M_1 & M_3 \\ M_2 & M_4 \end{bmatrix} \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}$$

In order to find stationary sources $\hat{s}^{\mathfrak{s}}(t)$ we only require that non-stationary contributions vanish, i.e. that $M_3 = 0$.

Symmetries and Invariances (I)

Let's express the true $A^{\mathfrak{s}}$ and $A^{\mathfrak{n}}$ as linear combinations of the respective estimated subspaces

$$\begin{aligned} A^{\mathfrak{s}} &= \hat{A}^{\mathfrak{s}} M_1 + \hat{A}^{\mathfrak{n}} M_2 \\ A^{\mathfrak{n}} &= \hat{A}^{\mathfrak{n}} M_4 \end{aligned}$$

The composite transformation (true mixing followed by the estimated demixing) reads

$$\begin{bmatrix} \hat{s}^{\mathfrak{s}}(t) \\ \hat{s}^{\mathfrak{n}}(t) \end{bmatrix} = \hat{B} A s(t) = \begin{bmatrix} \hat{B}^{\mathfrak{s}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{s}} A^{\mathfrak{n}} \\ \hat{B}^{\mathfrak{n}} A^{\mathfrak{s}} & \hat{B}^{\mathfrak{n}} A^{\mathfrak{n}} \end{bmatrix} s(t) = \begin{bmatrix} M_1 & 0 \\ M_2 & M_4 \end{bmatrix} \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}$$

In order to find stationary sources $\hat{s}^{\mathfrak{s}}(t)$ we only require that non-stationary contributions vanish, i.e. that $M_3 = 0$.

Symmetries and Invariances (II)

The separation of stationary and non-stationary sources is unique up to the linear transformations M_1, M_2 and M_4.

Symmetries and Invariances (II)

The separation of stationary and non-stationary sources is unique up to the linear transformations M_1, M_2 and M_4.

$$\begin{aligned} A^{\mathfrak{s}} &= \hat{A}^{\mathfrak{s}} M_1 + \hat{A}^{\mathfrak{n}} M_2 \\ A^{\mathfrak{n}} &= \hat{A}^{\mathfrak{n}} M_4 \end{aligned}$$

$\rightarrow$ we can identify the true non-stationary subspace but the not the true stationary subspace.

Symmetries and Invariances (II)

The separation of stationary and non-stationary sources is unique up to the linear transformations M_1, M_2 and M_4.

$$\begin{aligned} A^{\mathfrak{s}} &= \hat{A}^{\mathfrak{s}} M_1 + \hat{A}^{\mathfrak{n}} M_2 \\ A^{\mathfrak{n}} &= \hat{A}^{\mathfrak{n}} M_4 \end{aligned}$$

$\longrightarrow$ we can identify the true non-stationary subspace but the not the true stationary subspace.

$$\begin{bmatrix} \hat{s}^{\mathfrak{s}}(t) \\ \hat{s}^{\mathfrak{n}}(t) \end{bmatrix} = \begin{bmatrix} M_1 & 0 \\ M_2 & M_4 \end{bmatrix} \begin{bmatrix} s^{\mathfrak{s}}(t) \\ s^{\mathfrak{n}}(t) \end{bmatrix}$$

$\longrightarrow$ we can identify the true stationary sources (up to linear transformation) but not the true non-stationary sources.

Symmetries and Invariances

We are therefore free to choose the linear transformations M_1, M_2 and M_4 in order to formulate a more efficient algorithm.

Without loss of generality

We can center the data $x(t)$ and write the estimated demixing matrix as a whitening W followed by a rotation $\hat{B}^\top \hat{B} = I$:

$$\hat{A}^{-1} = \hat{B} W$$

with $W = \mathrm{Cov}(x)^{-1/2}$.

$\rightarrow$ The estimated stationary components then have zero mean and unit covariance.

Symmetries and Invariances

We are therefore free to choose the linear transformations M_1, M_2 and M_4 in order to formulate a more efficient algorithm.

Without loss of generality

We can center the data $x(t)$ and write the estimated demixing matrix as a whitening W followed by a rotation $\hat{B}^\top \hat{B} = I$:

$$\hat{A}^{-1} = \hat{B} W$$

with $W = \mathrm{Cov}(x)^{-1/2}$.

$\rightarrow$ The estimated stationary components then have zero mean and unit covariance.

Remaining task: Find orthogonal matrix $\hat{B}$ such that the estimated $\mathfrak{s}$-sources are *as stationary as possible.*

Optimizing for Stationarity

Working definition of stationarity

Divide the data into N epochs. Time series is stationary if its distribution is the same in each epoch.

Optimizing for Stationarity

Working definition of stationarity

Divide the data into N epochs. Time series is stationary if its distribution is the same in each epoch.

Optimization Problem

Find rotation $\hat{B}$ such that the estimated $\mathfrak{s}$-sources are as stationary as possible. Therefore: Minimize the difference in distribution between each epoch and the whole dataset.

Optimizing for Stationarity

Working definition of stationarity

Divide the data into N epochs. Time series is stationary if its distribution is the same in each epoch.

Optimization Problem

Find rotation $\hat{B}$ such that the estimated $\mathfrak{s}$-sources are as stationary as possible. Therefore: Minimize the difference in distribution between each epoch and the whole dataset.

Gaussian Approximation

Consider only differences in the first two moments of the distributions $\rightarrow$ Gaussian approximation (Max. Entropy Principle)

The Objective Function (I)

Objective function in words

Minimize the difference between the distribution of the estimated $\mathfrak{s}$-sources in each epoch and the whole dataset.

The Objective Function (I)

Objective function in words

Minimize the difference between the distribution of the estimated $\mathfrak{s}$-sources in each epoch and the whole dataset.

Distribution of the estimated $\mathfrak{s}$-sources in epoch i

Let $\hat{\mu}_i$ and $\hat{\Sigma}_i$ be the mean and covariance matrix of epoch i after centering and whitening. Given the rotation part $B^\top B = I$ of the estimated demixing, the mean and covariance of the estimated $\mathfrak{s}$-sources in epoch i can be written as

$$\hat{\boldsymbol{\mu}}_i^{\mathfrak{s}} = I^d B \hat{\boldsymbol{\mu}}_i \text{ and } \hat{\Sigma}_i^{\mathfrak{s}} = I^d B \hat{\Sigma}_i (I^d B)^\top$$

respectively, where I^d is the identity matrix truncated to the first d rows.

The Objective Function (II)

- We use the Kullback-Leibler divergence between Gaussians $\mathrm{KL}(\mathcal{N}(\mu_0, \Sigma_0) \,||\, \mathcal{N}(\mu_1, \Sigma_1))$ to measure differences in distribution

The Objective Function (II)

- We use the Kullback-Leibler divergence between Gaussians $\mathrm{KL}(\mathcal{N}(\mu_0, \Sigma_0) \,||\, \mathcal{N}(\mu_1, \Sigma_1))$ to measure differences in distribution
- Since we have centered and whitened the whole dataset, the estimated $\mathfrak{s}$-sources have zero mean and unit covariance.

The Objective Function (II)

- We use the Kullback-Leibler divergence between Gaussians $\mathrm{KL}(\mathcal{N}(\mu_0, \Sigma_0) \,||\, \mathcal{N}(\mu_1, \Sigma_1))$ to measure differences in distribution
- Since we have centered and whitened the whole dataset, the estimated $\mathfrak{s}$-sources have zero mean and unit covariance.

The objective function

$$\hat{B} = \operatorname*{argmin}_{B^\top B = I} \sum_{i=1}^{N} \mathrm{KL}\left[\mathcal{N}(\hat{\mu}_i^{\mathfrak{s}}, \hat{\Sigma}_i^{\mathfrak{s}}) \,||\, \mathcal{N}(0, I)\right]$$

$$= \operatorname*{argmin}_{B^\top B = I} \sum_{i=1}^{N} \left(-\log\det \hat{\Sigma}_i^{\mathfrak{s}} + \hat{\mu}_i^{\mathfrak{s}\top} \hat{\mu}_i^{\mathfrak{s}}\right)$$

Optimization in the Special Orthogonal Group (I)

The Optimization Problem

$$\hat{B} = \underset{B^\top B = I}{\operatorname{argmin}} \sum_{i=1}^{N} \left(-\log\det \hat{\Sigma}_i^{\mathfrak{s}} + \hat{\mu}_i^{\mathfrak{s}\top} \hat{\mu}_i^{\mathfrak{s}} \right)$$

In order to maintain the constraint $B^\top B = I$ we start with

$$\hat{B}^{\text{start}} = I$$

and find a multiplicative update $RR^\top = I$ in each step,

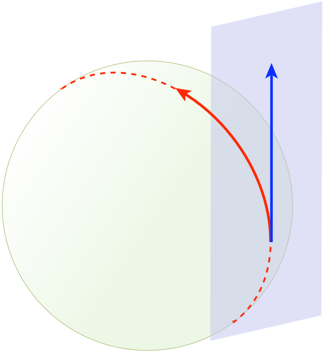
$$\hat{B}^{\text{new}} \leftarrow R\hat{B}.$$

Optimization in the Special Orthogonal Group (II)

We parametrize R as

$$R = \exp(M)$$

with $M^\top = -M$ and minimize w.r.t. M.



Optimization in the Special Orthogonal Group (II)

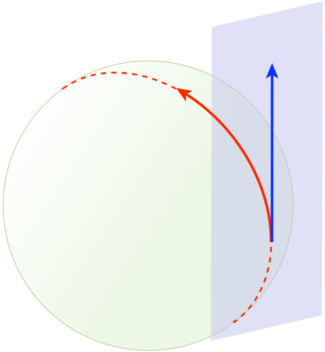
We parametrize R as

$$R = \exp(M)$$

with $M^\top = -M$ and minimize w.r.t. M.

Interpretation of elements M_{ij}:

Rotation angle of axis i towards axis j



Optimization in the Special Orthogonal Group (II)

We parametrize R as

$$R = \exp(M)$$

with $M^\top = -M$ and minimize w.r.t. M.

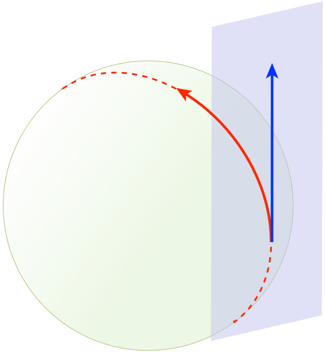
Interpretation of elements M_{ij}:

Rotation angle of axis i towards axis j

This leads to a gradient of the form

$$\left.\frac{\partial L_B}{\partial M}\right|_{M=0} = \begin{bmatrix} 0 & Z \\ -Z^\top & 0 \end{bmatrix}$$

where $Z \in \mathbb{R}^{d\times(D-d)}$.



Spurious Stationarity

- Directions in the non-stationary space can appear stationary if we have not observed enough variation.
- The presence of *spurious stationary directions* renders the true solution unidentifiable.

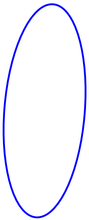
Spurious Stationarity

- Directions in the non-stationary space can appear stationary if we have not observed enough variation.
- The presence of *spurious stationary directions* renders the true solution unidentifiable.

How many distinct epochs do we need to rule out spurious stationary directions?

How many Epochs?

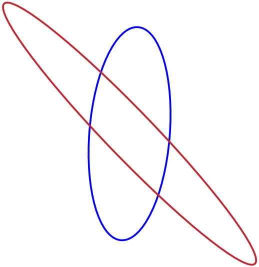
Imagine we have a two-dimensional non-stationary space and the epochs differ only w.r.t. the covariance matrix.



One epoch is certainly not enough.

How many Epochs?

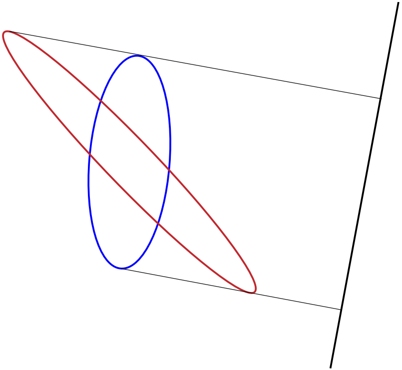
Imagine we have a two-dimensional non-stationary space and the epochs differ only w.r.t. the covariance matrix.



With two epochs there will in general exist a spurious stationary direction.

How many Epochs?

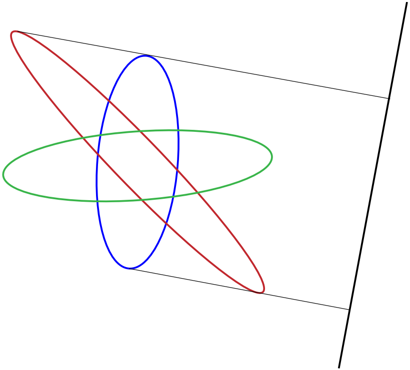
Imagine we have a two-dimensional non-stationary space and the epochs differ only w.r.t. the covariance matrix.



With two epochs there will in general exist a spurious stationary direction.

How many Epochs?

Imagine we have a two-dimensional non-stationary space and the epochs differ only w.r.t. the covariance matrix.



We need a third distinct epoch to eliminate this spurious stationary direction.

How many Epochs? – Theoretical Results

Theorem (Identifiability of SSA)

- *If the non-stationarity is expressed in* **both** *mean and covariances, the stationary subspace can be uniquely identified if*

$$N > \frac{D - d}{2} + 2.$$

How many Epochs? – Theoretical Results

Theorem (Identifiability of SSA)

- *If the non-stationarity is expressed in* **both** *mean and covariances, the stationary subspace can be uniquely identified if*

$$N > \frac{D-d}{2} + 2.$$

- *If the non-stationarity is* **only** *expressed in either mean* **or** *covariances, Identifiability is guaranteed for*

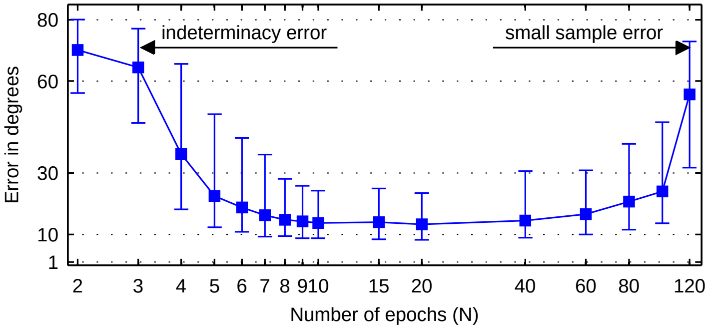
$$N > D - d + 1.$$

Simulations: Results

- 4 $\mathfrak{s}$-sources and 4 $\mathfrak{n}$-sources (correlated), random mixing matrix A.
- Fixed number of samples divided into epochs

Simulations: Results

- 4 $\mathfrak{s}$-sources and 4 $\mathfrak{n}$-sources (correlated), random mixing matrix A.
- Fixed number of samples divided into epochs



Brain-Computer-Interfacing (BCI)

- Control a device using intentional changes of the brain state (without muscles!)
- Brain states are detected using EEG-measurements
- Well discernable brain states: imagined (not executed!) movements of the left and right arm/feet.
- Berlin BCI: calibration phase / application phase

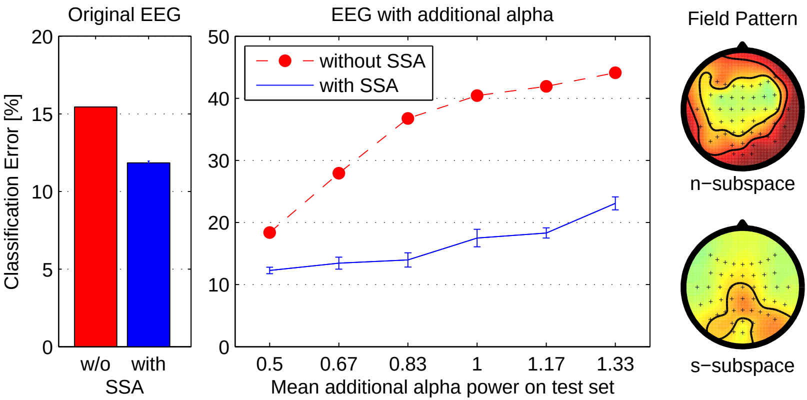
Brain-Computer-Interfacing (BCI)

- Control a device using intentional changes of the brain state (without muscles!)
- Brain states are detected using EEG-measurements
- Well discernable brain states: imagined (not executed!) movements of the left and right arm/feet.
- Berlin BCI: calibration phase / application phase

Problem

Classification accuracy deteriorates over time due to non-stationarities associated with fatigue (change in α-rhythm).

BCI Experiment: Results



Conclusion

SSA factorizes a time-series into stationary and non-stationary components.

- Two (mild) assumptions:
  - The non-stationarities affect the first two moments.
  - The observed time series is a linear superposition of stationary and non-stationary sources.
- The number of distinct epochs necessary scales *linearly* with the number of non-stationary sources.

Conclusion

SSA factorizes a time-series into stationary and non-stationary components.

- Two (mild) assumptions:
  - The non-stationarities affect the first two moments.
  - The observed time series is a linear superposition of stationary and non-stationary sources.
- The number of distinct epochs necessary scales *linearly* with the number of non-stationary sources.

Outlook

- Determining the number of stationary sources.
- Stationary w.r.t. time structure
- Efficient computation using methods from Algebraic Geometry
- Applications: finance, neuroscience, computer vision, etc.

Thank you very much.

Paul von Bünau, Frank C. Meinecke, Franz J. Király, Klaus-Robert Müller. Finding Stationary Subspaces in Multivariate Time Series. *Phys. Rev. Lett.* 103, 214101 (2009)