Supporting Information for the paper "Improving clustering by imposing network information"

Susanne Gerber and Illia Horenko
Institute of Computational Science, Università della Svizzera Italiana
Via G. Buffi 13,
6900 Lugano, Switzerland

June 1, 2015

This document contains the following sections:

1. Supplementary text
2. Captions for Movies SM1 to SM8
3. Supplementary Figures S1 and S2

1 Supplementary text

This section contains the following chapters:

- Relation between general parameter identification problems and clustering.
- Derivation of the graph-regularized clustering problem formulation.
- Derivation of the particular analytical solution for the Γ-optimization (Step ii) of the hard clustering algorithm ($\alpha = 1$)
- EEG data preprocessing pipeline.
- Classification of the clustering algorithms.

1.1 Relation between general parameter identification problems and clustering

Let $X = x(1), ...x(U)$ be the available measurement with $x_u = x(u)$ as an observation at a specific (time)-point or locations u (where u might relate to an index of a node/vertex on some graph $G = (E, V)$, $U = |V|$). Let us assume that the underlying dynamical process can be approximated by a mathematical direct model, $x_u = f(\theta(u))$. This expression is defined by the model function $f(\cdot)$ and a set of (a priori unknown) model parameters $\theta(u)$. This set of model parameters might be heterogeneous (i.e, explicitly u-dependent, which is opposed to the homogeneous/stationary model situation when this set of parameters is independent of u and $\theta(u) \equiv \text{const}$). When solving the inverse problem, we have to find an optimal set of model parameters $\theta^*(u)$ describing the dynamical process - given the observational dataset - in the "best" possible way. The expression "*best*" can become quantified in terms of a *fitness function* measuring the quality of the approximation by calculating the distance between the model's prediction and the original data set:

$$g(x_u, \theta(u)) : \mathbb{R}^n \times \Omega \mapsto \mathbb{R} \tag{1}$$

Then, optimal parameters $\theta^*(u)$ can be identified by a solution of the following variational problem:

$$l(\theta(u)) = \sum_{u=1}^{U} g(x_u, \theta(u)) \to \min_{\theta(u)}, \tag{2}$$

where l is referred to as the model fitness functional. Since the number of unknowns can be higher than the number of known parameters, the problem (2) is in general *ill-posed* and may thus lead to meaningless or unrobust solutions (when the optimal set of parameters $\theta^*(u)$ depends very sensitively on little changes in the data sequence X).
A way to overcome the ill-posedness of the variational problem is to assume the system's dynamics to be *locally stationary* with a time- or location- dependent switching process $\gamma_i(u)$, defined as the probability for the observation x_u to belong to the locally stationary model (or cluster) i, that is characterized by time independent model parameters θ_i.
This assumption is reasonable since in real world application - in most cases - the parameter function $\theta(u)$ changes much slower than the observable x_u itself.
This additional assumption can be formulated mathematically as a piece-wise linearity of the parameter function $\theta(u)$ in $\mathbf{K}$ clusters

$$\theta(u) = \sum_{i=1}^{\mathbf{K}} \gamma_i^{\alpha}(u)\theta_i \tag{3}$$

with $\alpha \geq 1$ being some scalar exponent (called fuzzifier) and cluster weights $\Gamma(u) = (\gamma_i(u), ...\gamma_K(u))$ being subject to the constraints

$$\sum_{i=1}^{\mathbf{K}} \gamma_i(u) = 1, \qquad \forall u \tag{4}$$

$$\gamma_i(u) \geq 0, \qquad \forall u, i = 1, ..., K. \tag{5}$$

Next, we insert (3) into (2). If one of the following is true: (i) if either the fitness function g from (2) is a metric wrt. θ, i.e. if it fulfills the triangle inequality and $g(x_u, \sum_{i=1}^{\mathbf{K}} \gamma_i^\alpha(u)\theta_i) \leq \sum_{i=1}^{\mathbf{K}} \gamma_i^\alpha(u) g(x_u, \theta_i)$; or (ii) if $\gamma_i(u)$ can only take the values 0 or 1, then the problem (3)-(5) has an exact upper bound denoted as the *average cluster functional* L :

$$L^\alpha(\Theta, \Gamma) = \sum_{u=1}^{T} \sum_{i=1}^{\mathbf{K}} \gamma_i^\alpha(u) g(x_u, \theta_i) \to \min_{\Theta, \Gamma}. \tag{6}$$

This functional is again subject to the constraints (4) and (5). In the case (i) $l(\theta(\cdot)) \leq L(\Theta, \Gamma)$, in the case (ii) $l(\theta(\cdot)) = L(\Theta, \Gamma)$.

1.2 Derivation of the graph-regularized clustering problem formulation

In order to tackle the still remaining issue of a possible ill-posedness, the variational problem needs a second step of regularization, confining some norm (e.g., the persistency/variation) of the unknown heterogeneous model parameter function $\theta(\cdot)$

$$\|\theta(\cdot)\|_G \leq \bar{C} \tag{7}$$

E.g., if the norm $\|\cdot\|_G$ is chosen as the total variation norm on a linear graph G representing the time axis with time moments of measurements to be its nodes (the particular case considered in the application example from the main manuscript text), this additional constraint (7) will bound the maximal number of transitions between the clusters by some (a priori unknown) constant $\bar{C}$. The idea standing behind exploits the observation that persistency is one of the main characteristic features of many real processes and that an appropriate mathematical regularization strategy is the clue to its efficient recovery from the observation data. It can be demonstrated [46, 16], that minimization of (6) with constraints (4),(5) and (7) is well posed and leads to a linear optimization problem with linear constraints.
If the additional persistency assumption (7) is excluded then the average cluster functional (6) can now be numerically minimized with almost every clustering method, thereby also providing the solution for the ill-posed heterogenous inverse problem (2). E.g., this can be done deploying the fuzzy-c-means algorithm where the cluster distance functional (1) takes the form of the square of the simple Euclidean distance between the points in n dimensions:

$$g(x_u, \theta_i) = \| x_u - \theta_i \|_2^2 . \tag{8}$$

If $\alpha = 1$, the resulting functional

$$L^\alpha(\Theta, \Gamma) = \sum_{i=1}^{\mathbf{K}} \sum_{u=1}^{T} \gamma_i^\alpha(u) \| x_u - \theta_i \|_2^2 \tag{9}$$

is identical to the classical k-means functional [25]. Also other classical methods of data analysis and machine learning (e.g., multilinear statistical regression, Gaussian mixture models (GMMs), and hidden Markov models (HMMs)) can be derived as special cases of the clustering problem (6), (4), (5) by

choosing specific model distance functions and regularity constraints. For details please see [16]. If the minimization of (6)-(5) is considered together with the additional regularity assumption (7) then the standard clustering algorithms are not applicable and the methods from the Finite Element Model family with Bounded Variation (FEM-BV) of model parameters should be taken [23, 24, 16].

The general weighted graph-based variation of the heterogeneous model parameters $\theta(u)$ on the graph can be defined as:

$$\|\theta(\cdot)\|_G = \sum_{u,v\in E} W_{u,v}||\theta(u) - \theta(v)|| \tag{10}$$

where $W_{u,v}$ is a matrix of kernel weights, e.g., $W_{u,v}$ is inverse-proportional to the distance between u and v on the grap G. Kernel weights can be additionally equipped with the probability measure $p(\cdot)$ for different edges, hereby including some probabilistic assumptions on the underlying dynamical process (e.g., Markovianity). This measure can, for example, refer to a probability $p(u)$ of finding a random Brownian walker - jumping between the connected nodes of a graph G for a very long time - at some specific node u. This measure is unique if the underlying Markov process on the graph is irreducible and aperiodic – in this case it can straightforwardly be computed, e.g., by calculating the dominant left eigenvector (i.e., the left eigenvector correspondent to the left eigenvalue 1.0) of the respective Markov transfer operator. In the situations when no a priori connection to the underlying Markov process is necessary (e.g., as it is the case in the results presented in a main manuscript text), this measure can be ignored by setting $p(\cdot) \equiv 1$. In these situations the underlying graph topology is solely reflected by the matrix of weights W.

Several possible ways of defining the kernel weight function and the probability measure exist, e.g., a very popular approach is to assume the graph to be correspondent to some Markov process, resulting in a so-called diffusion-distance $W_{u,v} = \exp[-\sigma \mathrm{dist}_G(u,v)]$ [47, 48, 49]. Following this idea, the graph-based 2-norm for the model parameters θ can be defined as:

$$\|\theta(\cdot)\|_G = \sum_{u,v\in E} W_{u,v} \left(\frac{\theta(v)}{p(v)} - \frac{\theta(u)}{p(u)}\right)^T \left(\frac{\theta(v)}{p(v)} - \frac{\theta(u)}{p(u)}\right) \tag{11}$$

The above expression can now be transformed by inserting the clustering assumption (3) into (11).

$$\|\theta(\cdot)\|_G = \sum_{u,v\in E} W_{u,v} \left(\sum_{i=1}^{\mathbf{K}} \left(\frac{\gamma_i^\alpha(v)}{p(v)} - \frac{\gamma_i^\alpha(u)}{p(u)}\right)\theta_i\right)^T \left(\sum_{i=1}^{\mathbf{K}} \left(\frac{\gamma_i^\alpha(v)}{p(v)} - \frac{\gamma_i^\alpha(u)}{p(u)}\right)\theta_i\right) \tag{12}$$

$$\leq \sum_{u,v\in E} W_{u,v} \sum_{i=1}^{\mathbf{K}} \left(\frac{\gamma_i^\alpha(v)}{p(v)} - \frac{\gamma_i^\alpha(u)}{p(u)}\right)^2 \|\theta_i\|^2 \tag{13}$$

Tikhonov-Regularization and smoothness of the cluster affiliation on the graph

Let $\widetilde{P}$ be in-degree weighted graph matrix (diagonal matrix containing the sum of edges weights terminating in u)

$$\widetilde{P}_{uu} \equiv \sum_{v|(v,u)\in E} W_{v,u} p^{-2}(u) \quad ; \quad \widetilde{P}_{uv(u\neq v)} \equiv 0 \tag{14}$$

and $\widetilde{Q}$ the out-degree weighted graph matrix (a diagonal matrix containing the sum of edges weights starting in u)

$$\widetilde{Q}_{uu} \equiv \sum_{v|(u,v)\in E} W_{u,v} p^{-2}(u) \quad ; \quad \widetilde{Q}_{uv(u\neq v)} \equiv 0 \tag{15}$$

Then from (11) and (13) it follows that

$$\|\theta(\cdot)\|_G \leq \sum_{i=1}^{\mathbf{K}} \|\theta_i\|^2 (\gamma_i^\alpha)^T D_G \gamma_i^\alpha \leq \overline{C}_{\mathbf{K}} < +\infty, \tag{16}$$

with $\overline{C}_{\mathbf{K}}$ being some a priori unknown finite constant, $D_G = \widetilde{P} - 2\widetilde{W} + \widetilde{Q}$ and $\widetilde{W}_{u,v} = \frac{W_{u,v}}{p(u)p(v)}$

Introducing a non-negative Lagrange-multiplier function ϵ^2, this constraint can be added as a penalty factor into the original clustering problem (6), resulting in the following Tikhonov-regularized optimisation problem

$$L^{\alpha,\epsilon}(\Theta,\Gamma) = \sum_{i=1}^{\mathbf{K}} \left[(\gamma_i^\alpha)^T g_i + \epsilon^2 \|\theta_i\|^2 (\gamma_i^\alpha)^T D_G \gamma_i^\alpha \right] \to \min_{\Theta,\Gamma} \tag{17}$$

subject to the constraints (4) and (5) with a vector of cluster affiliations $\{\gamma_i\}_u = \gamma_i(u)$, vector of cluster distances $\{g_i\}_u = g(x_u, \theta_i)$, $\sum_{i=1}^{\mathbf{K}} \gamma_i = (1_1, ..., 1_U) = 1^T$, and $\alpha \geq 1$. It is straightforward to verify that if the matrix D_G is positive-semidefinite then this regularized clustering problem represents an upper bound for the problems (6) and (1) (i.e., $\Lambda^\epsilon > L > l$). Solving this problem will be also providing approximate solutions of the problems (6) and (1).
The γ_i are here row vectors with U elements in each case, where $U = |V|$ is the number of all elements (vertices) of the graph. γ_i are vectors with the affiliations of each vertex to a cluster i. All probabilities of each vertex to belong to one of the clusters has to sum up to 1. Vector g contains the values of distances between the observation at a vertex u to the cluster i.
The minimization problem (17) can now be solved by means of a modification of the classical subspace iteration algorithm:

Clustering with an imposed graph information
Iteratively repeat until convergence in $L^{\epsilon,\alpha}$:

(Step i)	for a current value of Γ, (17) is minimized as an unconstrained convex (e.g., quadratic) problem wrt. Θ only (that can be done analytically, e.g., when g_i is an Euclidean distance);
(Step ii)	for a current value of Θ, (17) is minimized wrt. Γ, only as a constrained convex (e.g., as a quadratic programming - QP) problem.

1.3 Derivation of the particular analytical solution for the Γ-optimization (Step ii) of the hard clustering algorithm ($\alpha = 1$).

In order to solve (17) we first ignore the inequality-constraint $\gamma_i \geq 0$ (which would result in a quadratic minimization problem with equality and inequality constraints) and solve the problem only for equality constraints. We consider the most important case $\alpha = 1$ (which is the case for example for k-means clustering) and D_G being a symmetric matrix (which is the case if the underlying graph G is not a directed graph). Then deploying the method of Lagrange multipliers we obtain:

$$\forall_i : g_i + 2\epsilon^2 \|\theta_i\|^2 D_G \gamma_i + \lambda = 0. \tag{18}$$

This implies

$$\gamma_i = -\frac{1}{2\epsilon^2 \|\theta_i\|^2} D_G^{-1} (\lambda + g_i), \tag{19}$$

where D_G^{-1} denotes a (pseudo-)inversion operator for the graph distance matrix D_G, since D_G is positive (semi-definite) and might not be invertible in usual a sense. Since $\sum \gamma_i = 1$ this leads to

$$1 = -\frac{1}{2\epsilon^2\|\theta_i\|^2} D_G^{-1} \left(\mathbf{K}\lambda + \sum g_i \right) \tag{20}$$

$$1 + \frac{1}{2\epsilon^2\|\theta_i\|^2} D_G^{-1} \sum g_i = -\frac{\mathbf{K}}{2\epsilon^2} D_G^{-1}\lambda \tag{21}$$

$$\lambda = -\frac{2\epsilon^2\|\theta_i\|^2}{\mathbf{K}} D_G 1 - \frac{1}{\mathbf{K}} \sum_{i=1} g_i. \tag{22}$$

Inserting (22) in (19) we obtain

$$\gamma_i = -\frac{1}{2\epsilon^2\|\theta_i\|^2} D_G^{-1} \left(-\frac{2\epsilon^2\|\theta_i\|^2}{\mathbf{K}} D_G 1 - \frac{1}{\mathbf{K}} \sum_{i=1}^{\mathbf{K}} g_i + g_i \right) \tag{23}$$

$$= \frac{1}{\mathbf{K}} 1 + \frac{1}{2\epsilon^2 \mathbf{K}\|\theta_i\|^2} D_G^{-1} \sum_{i=1}^{\mathbf{K}} g_i - \frac{1}{2\epsilon^2\|\theta_i\|^2} D_G^{-1} g_i \tag{24}$$

$$= \frac{1}{\mathbf{K}} 1 - \frac{1}{2\epsilon^2 \mathbf{K}\|\theta_i\|^2} D_G^{-1} \left(-\sum_{i=1}^{\mathbf{K}} g_i + \mathbf{K} g_i \right). \tag{25}$$

With this expression for γ_i we get an explicitly-computable analytical solution. In the case of large ϵ this solution tends to a constant value ($\frac{1}{\mathbf{K}}1$) . Inequality constraint $\gamma \geq 0$ will be also preserved if D_G is posistive-semidefinite and if ϵ is chosen large enough.

This result is particularly important in context of the so-called Finite Element Method family of time series clustering methods with Bounded Variation of the model parameters (FEM-BV) [23, 24]. This method family represents a special case of the introduced graph-regularized clustering framework, when the underlying graph is linear (representing the time axis) with graph distances being localized only to consider/measure the nearest neighbour interactions in time. It is straightforward to verify that in this situation the graph distance matrix D_G will be tri-diagonal and positive-semidefinite, its inverse can be expressed analytically (through the eigenvectors and eigenfunctions of Laplace-operator in one dimension). This result allows to reduce the overall computational complexity of solving the (25) to $\mathcal{O}(U)$. Inverse of D_G in this tri-diagonal case can be analytically precomputed once and then re-used every time the (Step ii) of the clustering algorithm is performed. Therefore, the direct application of (25) for solving the γ_i-optimization substep of the FEM-BV-algorithms would allow to reduce their current complexity from $\mathcal{O}(U^p)$ (with $p > 2$) to $\mathcal{O}(U)$, thereby resolving the main current computational bottleneck of the FEM-BV-methods of time series analysis and allowing to address analysis of much longer time series then it is currently possible.

For small ϵ, the inequality constraint can not be guaranteed. In such situations the γ_i-optimization step of clustering algorithms needs to be solved as a quadratic minimization problem with equality and inequality-constraints, deploying the standard methods of sparse quadratic programming (QP) and initialising the iteration with the analytic solution (25).

1.4 EEG data preprocessing pipeline

The standard approach to analysis of EEG signals starts with defining the baseline as a normal state, or zero-line, respectively. During the data preprocessing step, the baselines are normally computed as empirical averages over the whole EEG time series. These averages are then subtracted from all other experiments. The resulting baseline-adjusted signals measuring the positive and negative deviations from the baseline are further analyzed. However, as known from the central limit theorem, empirical computation of the mean is subject to statistical uncertainty that is proportional to $\sigma/\sqrt{U}$, where σ is the variance of the signal and U is its length. Typically, noise-to-signal ratios are proportional to σ and for EEG signal they are of the order of thousands. This means that for very noisy signals with very large σ one would need to obtain very long EEG measurement series to be able to get a reliable estimate of the mean. On the other hand, the longer is the measured sequence, the higher is the probability that this baseline can not be approximated by a constant any more, since one can observe slow time-modulations of the baseline in longer experiments.
In order to overcome this drawback we proceeded as follows: not the EEG-signal as such, but the *differences* of the signal between the time steps

$$\Delta x_t^i = x_{t+1}^i - x_t^i, \tag{26}$$

were calculated from x^i (being the observed signal from the electrode $i = 1, ..., 64$ at time points $t = 1, .., U$). Herewith, we become independent from the definition of the baseline (and the respective statistical estimation error), since it is easy to verify that this transformation (26) is invariant wrt. any baseline shift of the original data x^i.
A second problem in the standard way of analyzing EEG data arises from the high number of dimensions - after the first preprocessing step the data matrix has a size of $\mathbb{R}^{64\times(T-1)}$. Since almost all clustering techniques that are potentially applicable to high-dimensional data are usually based on the Euclidean distance (2-norm) assumption, it is necessary to confirm the correctness of this assumption in every specific case. Therefore, the question we are going to answer within the next paragraph is: "How can the signal become preprocessed - without additional assumptions - such that it can be guaranteed that the Euclidean metric would become appropriate, even if it was not appropriate for the original data?"

Justification for using the 2-norm

A facility to make the Euclidean distance applicable is given by Hassler Whitney's the embedding theorem [50], stating that any smooth real m-dimensional manifold can be smoothly embedded in R^{2m+1}.
Furthermore, the embedding theorem of Takens [51], forming a bridge between nonlinear dynamical systems theory and the analysis of experimental time series can be utilized. This powerful theorem shows generically that a shadow-version of the original manifold can be reconstructed by analyzing it's time series projections via time-delay embedding. It is important to mention that if m is not known a priori, according to the Takens theorem it would be enough to obtain a reasonable lower bound of m since for all $m' > m$ the Takens-embedding will remain Euclidean in $2m' + 1$ dimensions. In the present problem - since m is not known a priori, we obtain its lower bound estimate by taking the delay-embedding of different length (i.e., considering different numbers of consecutive time instances of the full EEG signal as the one vector). Then, performing the Principal Component Analysis we can define m as the dimension of the essential linear manifold that contains over 90% of data variation. Following these procedures as formulated for the embedded versions of Principal Component Analysis [52, 53, 54, 55, 56, 57, 58], we obtained that for the delayed embeddings of at least 50 time instances the Euclidianity of the metric is provided for the analyzed EEG data series.

PCA analysis and data reduction

Both experiments were embedded with 50 dimensions and were conjointly analyzed via embedded versions of PCA [53, 54, 57]. Herewith, a mutual embedded PCA basis was defined with the help of which it is possible to examine both experiments with opened and closed eyes simultaneously. Following the PCA protocol, the joint covariance matrix (for embedded EEGs of opened and closed eyes measurements) was calculated as well as the dominant eigenvalues and eigenvectors were investigated.
This resulted in the finding that 300 dimensions are needed to reproduce over 90% of the original signal in both time series in this common embedded PCA basis.
Thus, we are now in the position to cluster/classify patterns that can be observed by the electrodes in a direct relation to the Euclidean dynamical system behind this measurements.

Scaling Invariance

Many clustering algorithms such as k-means are not scaling invariant and as such they are not suitable for solving the clustering problems where the data is represented in different units. A simple conversion of the units (e.g. from mV to V) would change the Euclidean distance and could result in different clustering results. This problem can be avoided by application of the Mahalanobis distance - introduced by P. Mahalanobis in 1936 [59]

$$D_M(x,\Theta) = \sqrt{(x-\Theta)^T \Sigma^{-1} (x-\Theta)} \tag{27}$$

that provides a relative measure for the similarity between the observation $x = (x_1, x_2, ..x_t)^T$ and a second data-set with cluster-center $\Theta = (\Theta_1, ..\Theta_2, ...\Theta_i)^T$ and covariance matrix Σ^{-1}. With this metric, the input data are transformed into a dataset in which all attributes have zero mean and unit variance. Herewith, the Maholanobis metric is invariant under any linear transformation of the original variables. This fact can be seen in the following way: if x (and thereby Θ) are expressed in units U, the covariance Σ will be expressed in units of U^2, its inverse Σ^{-1} will be expressed in units of U^{-2} and the whole expression in the right hand side of (54) will thereby become dimensionless/unitless. Due to this scaling invariance induced by the covariance distance, also the PCA-based manifold clustering deployed in the numerical example from this manuscript is scaling-invariant, simply because the linear manifolds Θ_i defining the clusters in this procedure are obtained from the dominant eigenvectors of the covariance matrices in the clusters, i.e., representing a reduced version of the scaling-invariant Maholanobis norm.
Summarizing the previous procedural steps we

- confirmed the correctness of using the 2-norm by constructing an appropriate time-delayed embedding of the original data;
- are able to reconstruct the attractor characterizing the whole dynamical system (getting use of the Takens theorem that states that the observed output is confined to an attractive manifold characterizing the dynamical system in Euclidean space);
- created a mutual (shift-invariant) basis for the two independent experiments;
- ensured by utilizing the PCA-based metric $g(x_u, \Theta_i) = \|x_u - \Theta_i \Theta_i^T x_u\|$ (being a reduced version of the Mahalanobis metric) that our treatment can be considered as scaling-invariant.

1.5 Classification of the clustering algorithms

In the following, we give a brief classification of the clustering algorithms for which the presented methodology of imposing the graph/network-related a priori information can be deployed.
Clustering algorithms can be classified according to the type of data-input to the algorithm, the clustering criterion defining the similarity between data points and the fundamental (theoretical) concepts on which clustering analysis techniques are based.
This categorization is neither unique nor canonical and there exist a multitude of different approaches for classification of clustering approaches [1, 2, 3, 4].

- **Partitional (or centroid-based) clustering.** This approach, with $k-$means [60, 61] being the most popular representative of this method family, attempts to directly decompose a data set with U objects into k disjoint clusters such that the partitions optimize a certain criterion. Each cluster is represented by a centroid (the "center of gravity") which may not necessarily be an object from the data set. Typically, k seeds are randomly selected and a relocation scheme iteratively reassigns points between the clusters to optimize the clustering criterion. The minimization of the square-error criterion (the sum of squared Euclidean distances of points from their closest cluster centroid) is the most commonly used setting for k-means. Numerous improvements and variants of k-means clustering have been introduced such as *k-medoids* [62], *k-medians clustering* [63] and *K-means++* [64].

- **Hierarchical (or connectivity-based) clustering [65]:** This family of approaches proceeds by either merging smaller clusters at each step into larger ones (agglomerative algorithms) or by splitting the data repeatedly into finer groups (divisive methods). The result of a hierarchical clustering is a dendrogram which is a tree of clusters, with a single all-inclusive cluster at the top and singleton clusters of individual points at the bottom. Prominent representatives of hierarchical clustering algorithms are BIRCH [66], CURE [67], and CHAMELEON [68].

- **Model-based clustering:** comprises also probabilistic- or distribution-based clustering and contains approaches which use certain (probabilistic) models for clusters, attempting to optimize the fit (e.g., maximizing the respective log-likelihood) of the data through the model. The most widely used model-based clustering methods are Bayesian approaches, e.g., Hidden Markov Models [69, 70] the Gaussian Mixture model (GMM) approach [71, 72] based on the EM (Expectation-Maximization) algorithm [73] and the "moving windows methods" [74, 75]. Other model- based approaches are e.g neural network approaches with SOM (self-organizing feature map) [76].

- **Density-based clustering** is based on the key-idea of grouping neighboring data points into clusters based on density conditions as the local cluster criterion. The major features are the discovery of clusters of arbitrary shape and the handling of noisy data. Widely known representatives in this group are DBSCAN [77, 78] , OPTICS [79], CLIQUE [80] or the very recent approach from Rodriguez and Laio [81]. These algorithms are less sensitive to outliers and can discover clusters of irregular shapes. However, their applicability is typically confined to the low-dimensional data.

- **Spectral clustering** computes a similarity matrix between all pairs of data points. An eigenvalue decomposition is then performed, data points are projected into a space spanned by a subset of the eigenvectors and one of the root-algorithms (typically k-means or hierarchical clustering) is used to cluster the data [82].

- **Grid-based clustering:** Recently, a number of clustering algorithms for spatial data have been developed. These algorithms quantise the space into a finite number of cells and then do all operations on this quantised space. Representatives of this category are STING (Statistical Information Grid-based method) [83] and WaveCluster [84].

- **Soft- (or overlapping) clustering:** All algorithms described above result in non-overlapping groups, so-called hard cluster, where each data-object is grouped in an exclusive way, and belongs to exactly one cluster. Moreover, all points classified into the same cluster belong to it with the same degree of belief (i.e., all values are treated equally in the clustering process). The issue of uncertainty support in clustering tasks had lead to the introduction of approaches that use fuzzy logic concepts in their procedure [14]. One of the most prominent fuzzy clustering algorithms is the fuzzy c-Means (FCM) [15]. FCM attempts to cluster data such that each data-object may belong to several clusters with different degrees of membership.

- **FEM-BV:** The **F**inite **E**lement time series analysis **M**ethodology with **B**ounded **V**ariation of model parameters (FEM-BV) [47, 48] deployes the idea of temporal BV-regularization in the context of time series clustering problems. Combining the computational ideas like adaptive Finite Element Methods from the area of partial differential equations (PDEs) and regularization of cluster affiliations functions with the distance metrics deployed in other clustering methods, it allows to make the algorithms (from the other method families described above) more robust and less ill-posed. The distinctive property of FEM-BV is, that the methodology generalizes most of the standard parametrical data-analysis methods, e.g., regularized regression/spline methods [50, 45], multivariate autoregressive models with external factors (VARX), discrete homogenous Markov and Bernoulli processes, and principal component analysis (PCA), Smoluchowki and Langevin stochastic dynamics and the above mentioned clustering methods into a unified, non-parametric and non-stationary setting (please see [16] for examples of these generalizations). Thereby, all of these different methods can be handled numerically in context of the same theoretical and algorithmic framework, borrowing a lot of components and concepts from the area of partial differential equations and standard numerical optimization methods. In combination with information theory, the FEM-BV-framework allows an adaptive data-based inference of the most appropriate dynamical model that, from the viewpoint of the information, describes the time series data in an optimal way.

2 Supplemental Movies Legends

Supporting Movie SM1 Visualization of the most dominant attractor dimension over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with opened eyes.

Supporting Movie SM2 Visualization of the most dominant attractor dimension over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with closed eyes.

Supporting Movie SM3 Visualization of fourth eigenvector over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with opened eyes.

Supporting Movie SM4 Visualization of the fourth eigenvector over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with closed eyes.

Supporting Movie SM5 Visualization of the seventh eigenvector over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with opened eyes.

Supporting Movie SM6 Visualization of the seventh eigenvector over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with closed eyes.

Supporting Movie SM7 Visualization of the tenth eigenvector over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with opened eyes.

Supporting Movie SM8 Visualization of the tenth eigenvector over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system) for the experiment with closed eyes.

3 Supplementary Figures S1 and S2

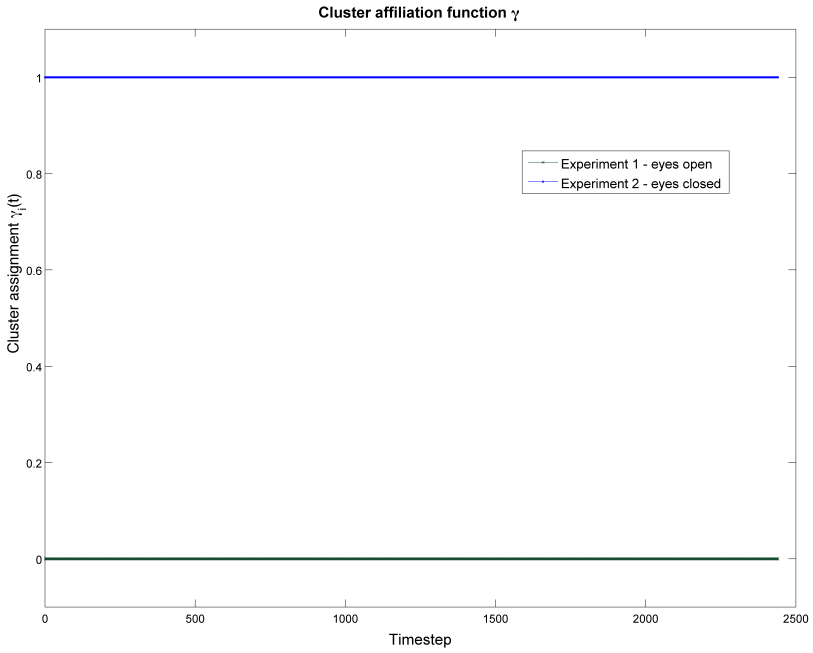

Figure S 1: Cluster affiliation function for the two identified manifolds

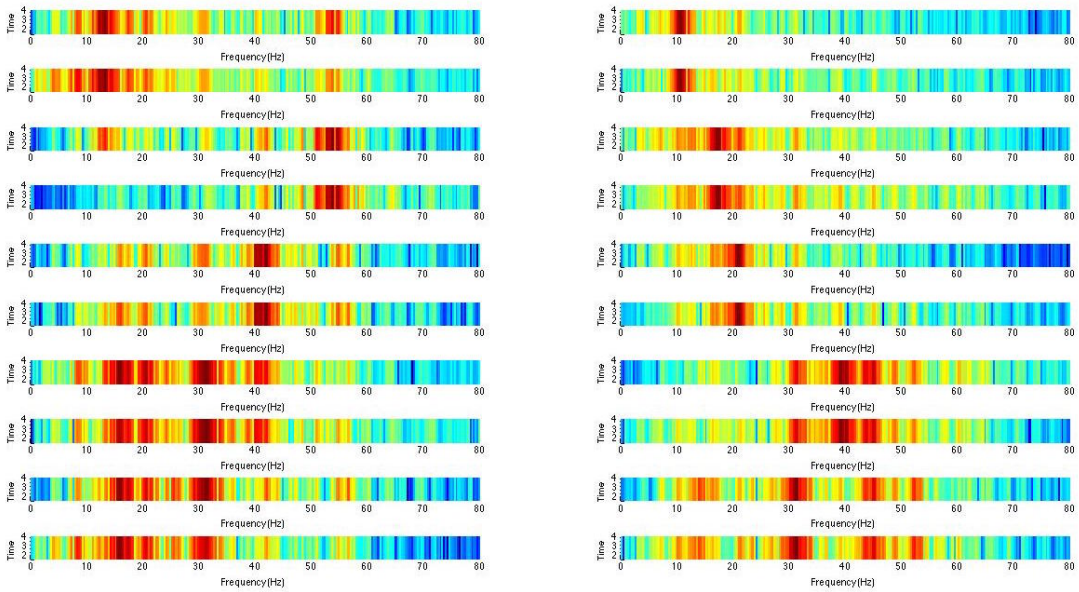

Figure S 2: Spectrograms of the EEG data. The left column shows the spectrogram of the projected experiment "eyes open" and the right column the spectrogram of projected experiment "eyes closed". Each row is computed from the projection of the EEG data on one of the first 10 dominant dimensions of the attractor, starting with the main component in the first row. Herewith, the periodic patterns holding the biggest proportion of the dynamical systems behavior in the two identified cluster states is identified and visualized. The two experiments show different oscillation patterns, especially concerning the first six main components (upper six rows of the Figure). The first two main signal components (in first two rows of the Figure) of the experiment with open eyes are bands with frequencies of about [12-25 Hz], which is exactly the frequency range for rhythmical beta activity (low beta waves [12.5-16 Hz] and medium beta waves [16.5-20 Hz]). Furthermore, in both first two main components - and even stronger in components 4-8 (raws 4 to 8 in the Figure) - patterns of neural oscillation with frequencies in the range of [40-60 Hz] are extracted. These waves are most likely gamma waves. Whereas in components 3-6 the predominant patterns are these gamma waves, in components 7 to 10 the dominant dynamics is represented by a mixture/combination of alpha-, beta- and gammarhythms as well as - most likely - of further brain waves such as SMR, theta and/or mu waves. The main two manifold components of the EEG signal for closed eyes are clearly defined bands with frequencies of about [8-13 Hz] which are the typical alpha rhythms (possibly with an admixture of mu rhythm). The main patterns in components 3-6 in the EEG with closed eyes can be clearly dedicated to beta rhythm activity, components 7 and 8 are obviously affected by gamma waves and components 9-10 are mixtures of alpha- beta - and gamma-waves.