APPLIED MATHEMATICS

Improving clustering by imposing network information

10.1126/sciadv.1500163

Susanne Gerber and Illia Horenko*

Cluster analysis is one of the most popular data analysis tools in a wide range of applied disciplines. We propose and justify a computationally efficient and straightforward-to-implement way of imposing the available information from networks/graphs (a priori available in many application areas) on a broad family of clustering methods. The introduced approach is illustrated on the problem of a noninvasive unsupervised brain signal classification. This task is faced with several challenging difficulties such as nonstationary noisy signals and a small sample size, combined with a high-dimensional feature space and huge noise-to-signal ratios. Applying this approach results in an exact unsupervised classification of very short signals, opening new possibilities for clustering methods in the area of a noninvasive brain-computer interface.

INTRODUCTION

Clustering can be considered as one of the most important learning and data analysis methods in a large variety of applied disciplines. The general aim of clustering is in assigning a set of objects into groups, such that the degree of association—the "similarity"—becomes maximized between members of the same group/cluster and minimized between members of different groups/clusters. The concept of similarity is—in most cases—realized via a distance function on the data space. A wide variety of different algorithms have been developed for special topics, each of them based on a diverse notion of what constitutes a cluster, thus using different induction principles for the clustering procedures. Consequently, the application of even slightly different algorithms may result in very deviating outcomes, even applied to the same data under the same conditions (*1*–*4*). Because the choice of "the best possible" clustering algorithm highly depends on the individual data set and the intended use of the results, clustering algorithms are developing just as dynamically as technical capabilities are evolving. In particular, the new possibilities offered by high-performance computing technologies give rise to exciting advances in the ability of researchers to use cluster modeling to analyze huge and complete data sets, from climate/weather research (*5*, *6*), economics and finance (*7*, *8*), computational biology, biophysics/bioinformatics (*9*, *10*), and neuroscience (*11*, *12*).

Most often in data analysis, the available information is only a series of U (experimental) observations $X = x(1), \ldots, x(U)$, without a detailed knowledge about the underlying model $x(u) = f(\theta(u))$ and its parameters $\theta(u)$. Inference of the "optimal" parameters in a wide range of data analysis approaches is mostly achieved by minimizing the appropriate fitness function:

$$l(\theta(u)) = \sum_{u=1}^{U} g(x(u), \theta(u)) \quad (1)$$

It is measuring the quality of the model for describing the given data sequence X by calculating the sum of distances $g(x(u), \theta(u))$ between the model's prediction [obtained from the parameter values $\theta(u)$] and the analyzed data $x(u)$. However, if only one data sequence X is available and the parameter θ is allowed to change with u in an arbitrary way, the respective minimization problem will simply not have enough data X to estimate all of the values of $\theta(u)$ and will result in what is called "overfitting."

Assuming $\theta(u) = \sum_{i=1}^{K} \gamma_i(u)\theta_i$ (with $\sum_{i=1}^{K} \gamma_i = 1$ and $\gamma_i(u) \geq 0$ for all i and u), clustering algorithms can also be formulated and implemented as a minimization with respect to (wrt.) $\Theta = (\theta_1, \ldots, \theta_K)$ and $\Gamma = (\gamma_1, \ldots, \gamma_K)$ of the so-called clustering functional L^{α}, taking the form (*13*, *14*):

$$L^{\alpha}(\Theta, \Gamma) = \sum_{i=1}^{\mathbf{K}} (\gamma_i^{\alpha})^T g_i \quad (2)$$

where T is a vector transposition operation, $\{g_i\}_u = g(x(u), \theta_i)$ are the row vectors of cluster distances, $\{\gamma_i\}_u = \gamma_i(u)$ are the row vectors of cluster affiliations [that is, $\gamma_i(u)$ is the probability for data point u to be from cluster i], and $\alpha \geq 1$ is a fixed scalar exponent called the fuzzyfier (*14*, *15*), allowing for the assignment of observed data points to more than one cluster (soft clustering) if $\alpha > 1$.

It can be demonstrated (please see section 1 in the Supplementary Text for a detailed proof) that $L^{\alpha} \geq l$, that is, that a wide range of clustering algorithms in their minimizational formulation (Eq. 2) represent methods for minimizing the upper bound L^{α} of the more general fitness function l. This means that by doing clustering, one is implicitly finding an approximate piecewise homogeneous solution of the more general and heterogeneous (that is, with parameters θ being dependent on u) data analysis problem (Eq. 1).

Most of the existent clustering methods can be formulated as the optimization of Eq. 2, differing only in the choices of α and the distance function $g(x(u), \theta_i)$. If, for example, $\alpha = 1$ and $g(x(u), \theta_i) = ||x(u) - \theta_i||_2$, then minimization of Eq. 2 is the task performed by the classical k-means algorithm (*16*). Also, other classical methods of data analysis and machine learning (for example, multilinear statistical regression, Gaussian mixture models, and hidden Markov models) can be derived as special cases of the above optimizational formulation by choosing specific model distance functions and additional constraints (*16*). In other words, clustering can be treated as an inverse problem that can become expressed and solved as a minimization—or a maximization—problem.

Università della Svizzera Italiana, Via Giuseppe Buffi 13, 6900 Lugano, Switzerland
*Corresponding author. E-mail: horenkoi@usi.ch

Computationally, the procedure of finding a minimum of L^{α} wrt. Θ and Γ in all of the clustering algorithms is implemented iteratively, by a consecutive repetition of the two following steps:

Standard Clustering
Iteratively repeat until convergence in L^{α}:
(Step i) for a current value of Γ, Eq. 2 is minimized to wrt. Θ only (that can be done analytically, for example, in the case of classical *k*-means);
(Step ii) for a current value of Θ, Eq. 2 is minimized to wrt. Γ only (that can also be done analytically).

It is straightforward to verify that for any initial choice of Γ or Θ, iterative repetition of these two steps will lead to a monotonic minimization of L^{α} and convergence to some local minimum of L^{α} (*13*). The obtained final estimates of Γ and Θ for different local minima can then be compared with respect to their values of L^{α} to identify the most optimal clustering of the data sequence X - corresponding to the smallest value of the respective clustering functional from Eq. 2.

One of the central problems of virtually all state-of-the-art clustering approaches lies in the fact that the related optimization problem is generally nonconvex (may have various local minima) and nonrobust or, in mathematical terms, ill-posed, meaning that tiny changes in the start conditions or in the tuning parameters of the algorithm might result in large changes in the answer. As mentioned above, the latter issue may arise due to the high number of unknowns in relation to the known parameters. Therefore, the results may heavily depend on the choice of the initialization, or tuning parameter sets, thus involving the risk of overfitting and frequently making the clustering results nonreproducible even on the same computer with the same set of user-defined tuning parameters (*16*).

This drawback becomes even more apparent when dealing with high-dimensional data, where most of the existing methods can fail due to the various problems of popular distance metrics deployed in clustering algorithms. To overcome the above-mentioned issue of nonrobustness, a frequently deployed strategy is to add some additional information or assumptions to the problem. In mathematical terms, this strategy is called regularization. Herewith, the nonrobust problem can be transformed into a robust one (*17*). Tikhonov regularization (*18*) and LASSO regularization (*19*) are prominent examples in the context of, for example, spline interpolation and parametric regression problems in data analysis and statistics.

MATERIALS AND METHODS

The central methodological contribution of this manuscript is in finding and justifying a computationally efficient and easy-to-implement way of imposing additional information—given in the form of a graph or a network—on clustering algorithms.

The main idea is based on the insight that the data index u can be brought in a relation to the respective node in some network or graph $G = (E, V)$, with edges E and vertices V, $U = |V|$. This graph G is equipped with some graph-related variation measure $||\theta(\cdot)||_G$, for example, with $\|\theta(\cdot)\|_G = \sum_{u,v\in E} W_{u,v}\|\theta(u) - \theta(v)\|_2^2$ squared Euclidean norm of the parameter θ variation on the graph G. In this expression, W is a matrix of weights, with elements $W_{u,v}$ being, for example, inverse-proportional to the Euclidean length of a minimal path connecting vertices u and v on the graph G. This graph G and the variation measure $||\cdot||_G$ are assumed to be underlying the measurement process and to be known a priori.

Then, inserting the clustering assumption $\theta(u) = \sum_{i=1}^{K}\gamma_i(u)\theta_i$ from above into $||\theta(\cdot)||_G$, we obtain

$$\|\theta(\cdot)\|_G \le \sum_{i=1}^{K}\|\theta_i\|^2(\gamma_i^{\alpha})^T D_G\gamma_i^{\alpha} \le \overline{C}_{\mathbf{K}} < +\infty, \qquad (3)$$

where $\overline{C}_{\mathbf{K}}$ is some (unknown) constant, $||\theta_i||$ is a Euclidean norm for cluster parameters θ_i and $D_G = P - 2W + Q$ (with diagonal matrices $P_{uu} \equiv \sum_{v|(v,u)\in E} W_{v,u}$ and $Q_{uu} \equiv \sum_{v|(u,v)\in E} W_{u,v}$; please see chapter 1.2 of the Supplementary Text for a detailed derivation). To give a concrete example, when dealing with problems of time series analysis, index u is denoting the time index of every particular data point, and the underlying graph G is a linear graph shown in Fig. 1A. Then, kernel weight W can be defined as $W_{u,v} = 1$ (for $||u - v|| = 1$) and $W_{u,v} = 0$ (for $||u - v|| \neq 1$), and the resulting D_G will be a tridiagonal positive semidefinite symmetric Laplacian matrix. This case will be particularly important in a context of time series clustering methods considered below.

Deploying Eq. 3 as an additional constraint in the minimization of the original clustering problem (Eq. 2), one gets a possibility of a "guided" search for parameters $\theta(\cdot)$—based on their differences measured in terms of the a priori available graph/network information G. More specifically, choosing large values of $\overline{C}_{\mathbf{K}}$ will result in those possible solutions of the original clustering problem that are very different wrt. parameters θ on the neighboring nodes of the graph G. Decreasing the value of $\overline{C}_{\mathbf{K}}$ will provide parameters that are more and more "close" in terms of the a priori available information G. Finally, setting $\overline{C}_{\mathbf{K}}$ to zero will result in stationary/homogeneous estimates, that is, in the parameters θ that are equal for all of the graph G nodes and, thereby, for all of the data points u. Formulated as an additional assumption, Eq. 3 basically means that we expect any two different nodes u and v of the graph G (that are not "too far away" from each other in a sense of the underlying graph distance measure) to have not "too different" values of unknown parameters $\theta(u)$ and $\theta(v)$.

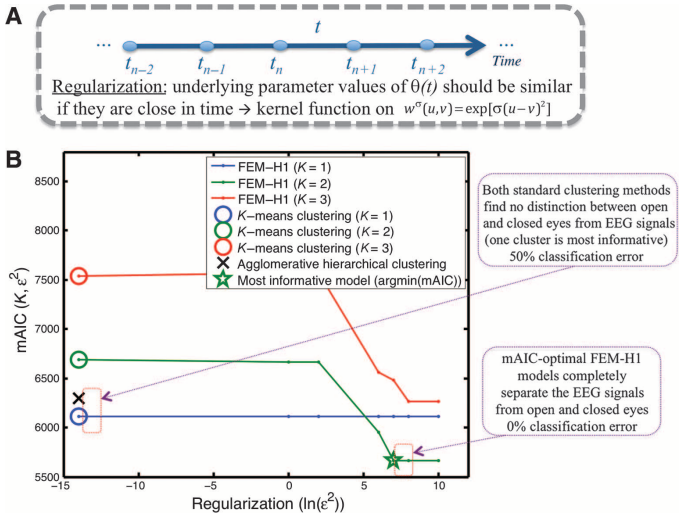

Fig. 1. An example of the imposed network and a cluster model discrimination. (**A**) Imposed (linear) graph: a priori persistency assumption for the underlying dynamics in time. (**B**) Comparing information content of EEG clusterings: graphs of the AIC values for $K = 1$ to 3 as a function of the regularization constant ϵ^2.

It turns out that from numerical/computational points of view, instead of directly deploying Eq. 3 as an additional inequality constraint, it is more advantageous to use its equivalent formulation—when the confined term from Eq. 3 is added as a penalty term to the original clustering problem formulation (Eq. 2):

$$L^{\epsilon,\alpha}(\Theta,\Gamma)=\sum_{i=1}^{\mathbf{K}}[(\gamma_i^{\alpha})^T g_i+\epsilon^2||\theta_i||^2(\gamma_i^{\alpha})^T D_G(\gamma_i^{\alpha})], \quad (4)$$

and subject to the original constraints $\sum_{i=1}^{K}\gamma_i=(1_1,\ldots,1_U)=1,\gamma_i\geq 0$. Penalty factor ϵ^2 can then be interpreted as the level of our confidence in the a priori network/graph information G: if ϵ^2 is large, then the impact of the a priori information on the overall solution will be strong; if ϵ^2 is very small or zero, then we essentially solve the original clustering problem (Eq. 2) and information about G does not play any significant role. Systematic strategies for automated and user-independent choice of ϵ^2 and $\mathbf{K}$ will be introduced below.

The main theoretical/conceptual advantage of this problem formulation (Eq. 4) can be revealed after making an observation that the right-hand side of the obtained functional is essentially a log-likelihood formulation of the Bayes theorem: with the second term being a log-likelihood of the prior distribution in the space of clustering parameters (that becomes Gaussian if one fixes either Γ or Θ) and with the first term keeping the form of the original clustering functional (Eq. 2) as a posterior log-likelihood [It is interesting to observe that the direct application of Bayes theorem to the original clustering problem (Eq. 2), assuming the independence of Γ or Θ and a priori Gaussianity for both of them, would result in two separate quadratic terms for Θ and for Γ in the right-hand side of Eq. 4, that is, in $\epsilon_\Gamma^2\gamma_i^T D_\Gamma\gamma_i+\epsilon_\Theta^2\theta_i^T D_\Theta\theta_i$. In contrast, we obtain a fourth-order log-likelihood term that is derived from an assumption (Eq. 3) and does not require or impose a priori conditional independence of Γ and Θ.]. In another words, increasing the value of the user-defined tunable parameter ϵ^2 would decrease the variance/uncertainty of cluster parameters Γ or Θ, making this transformed clustering problem well posed and robust, that is, less and less dependent on small changes of algorithmic parameters and on fluctuations in the analyzed data X.

Moreover, it is straightforward to validate (please see chapter 2 of the Supplementary Text for the detailed derivation) that this regularized clustering problem represents an upper bound for problems 1 and 2.

$$l(\theta(\cdot))\leq L^{\alpha}(\Theta,\Gamma)\leq L^{\epsilon,\alpha}(\Theta,\Gamma), \quad (5)$$

that is, the ill-posed data analysis problem (Eq. 1) (and the clustering problem in Eq. 2) can be both approximately solved through a minimization of the well-posed problem (Eq. 4).

The transformed problem (Eq. 4) can now be solved by a slight modification of the classical clustering algorithm explained above:

Clustering with an imposed graph information
Iteratively repeat until convergence in $L^{\epsilon,\alpha}$:
(Step i) for a current value of Γ, Eq. 4 is minimized as an unconstrained convex (for example, quadratic) problem wrt. Θ only (this can be done analytically in many cases);
(Step ii) for a current value of Θ, Eq. 4 is minimized to wrt. Γ only, as a constrained convex [for example, as a quadratic programming (QP)] problem.

It is easy to prove that the iterative repetition of these two convex minimization steps will monotonically converge to a local minimum of $L^{\epsilon,\alpha}$.

Step ii of this modified clustering algorithm represents the main computational bottleneck because it requires a numerical solution of a large QP problem. If the a priori graph/network information is available in the form of an undirected graph G, the corresponding matrices D_G will be symmetric. If D_G are also sparse, then we can take advantage of the highly efficient linear solvers for sparse banded matrices implemented, for example, in high-performance software libraries such as (Sca)LAPAK (*20*, *21*) or FLAME (*22*). However, the computational complexity in this case will still scale as $\mathcal{O}(U^p)$ (with $p>2$), limiting the overall applicability of this methodology to relatively small data sets.

By deploying the method of Lagrange multipliers, it can be shown that in a particular case of hard clustering (that is, when $\alpha=1$), one can find an explicit analytical solution of this problem:

$$\gamma_i=\frac{1}{\mathbf{K}}1-\frac{1}{2\epsilon^2\mathbf{K}||\theta_i||^2}D_G^{-1}\left(\mathbf{K}g_i-\sum_{i=1}^{\mathbf{K}}g_i\right), \quad (6)$$

that also satisfies the optimization constraints $\gamma_i\geq 0$ and $\Sigma_i\,\gamma_i=1$. Please see chapter 3 of the Supplementary Text for a detailed derivation of this solution.

This result (Eq. 6) is particularly important in the 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 only being localized to consider/measure the nearest neighbor interactions in time. In this situation, the graph distance matrix D_G will be a Laplacian matrix, that is, it will be tridiagonal and positive-semidefinite, meaning that 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 step ii in FEM-BV clustering methods from $\mathcal{O}(U^p)$ (with $p>2$) to $\mathcal{O}(U)$. The inverse of D_G in this tridiagonal case can be analytically precomputed once and then reused with Eq. 6 every time when step ii of the FEM-BV clustering is performed. This resolves the main current computational bottleneck of the FEM-BV methods of the time series analysis, allowing us to address the analysis of a much longer time series than what is currently possible with these methods.

Choosing an optimal setting for the clustering algorithm

Because the outcome of clustering is highly dependent on the specific choice of the number of clusters $\mathbf{K}$, regularization constant ϵ^2, and fuzzier α (as well as on the choice of the distance measure on the graph G), another challenge is in choosing all of these parameters and settings in some "optimal" way. Several ways of choosing optimal clustering parameters have been discussed in the literature (*23*, *24*). In the situations when the distance function g can be interpreted in the probabilistic sense—for example, as a log-likelihood of some distribution—information criteria can be used to identify/discriminate the optimal parameter setting for clustering algorithms (*16*, *25*). This is done in a sense of "Occam's razor" principle: information-theoretic tools such as Akaike (AIC) and Bayes information criteria can be used to find the clusterings that are simultaneously most qualitative (in terms of minimizing the value of L

in Eq. 2 or 4) and simple (in terms of the low number of clusters and/or other tunable parameters). However, this information-theoretic approach is limited to situations when g has an explicit probabilistic interpretation, that is, has a form of the parametric (for example, Gaussian) log-likelihood. It is not the case in situations when g simply measures some geometric distance (for example, an Euclidean distance between the points for standard k-means clustering) that is not a priori-interpretable as some log-likelihood associated with some assumed parametric probability measure (for example, Gaussian). In (*16*), it was presented how to select the optimal clustering models in a non-parametric way, without a priori parametric probabilistic assumptions on underlying measures. As was demonstrated in information theory (*26*), nonparametric exponential distributions represent the family of maximum entropy distributions for the given scalar valued process time series. That is, max log-likelihood fitting of these exponential distributions to the available data would result in the posterior identification of the most likely and least-biased (that is, the simplest) random process that has the highest affinity to produce such an output as the one that is observed. The main idea of the respective nonparametric model selection approach is based on the posterior maximum log-likelihood fitting of a sequence of such nonparametric exponential family distributions

$$\rho(y(u;\epsilon,\mathbf{K},\alpha);\lambda^{\epsilon,\mathbf{K},\alpha,k}) = \exp\left(\sum_{j=0}^{k}\lambda_j^{\epsilon,\mathbf{K},\alpha,k}y^j(u;\epsilon,\mathbf{K},\alpha)\right), \qquad (7)$$

(for various k) to the scalar sequence of the posterior clustering model errors $y^{\epsilon,\mathbf{K},\alpha}(u)=\sum_{i=1}^{K}\gamma_i(u)g(x(u),\theta_i)$, obtained from regularized clustering problem (Eq. 4) for various combinations of ϵ, $\mathbf{K}$, α. Then, the log-likelihood of the identified nonparametric process (Eq. 7) can be plugged into the information criterion to obtain the most informative model. This modified AIC (mAIC) [please see section 3 of (*16*) for more details] will be, in the following text, applied to compare different clustering results and to identify the optimal combination of clustering parameters $\mathbf{K}$, k, ϵ^2, and α, resulting in the procedure that is free of any user-defined tunable parameters.

Application example

As explained in Materials and Methods, the FEM-BV family of time series clustering methods represents a particular case of the introduced graph-regularized clustering methodology, with the graph G being a linear graph and matrix D_G being tridiagonal. Application of the methodology in this situation to various test systems, comparison to standard clustering methods, and machine learning approaches can be found, for example, in the review paper (*16*).

In the following text, we illustrate the potential of the presented methodology for time series analysis (that is, imposing the linear graph G as in the case of FEM-BV methods) and exploiting the new possibilities provided by the formulation (Eq. 4) and, especially, by the obtained analytical solution (Eq. 6). We consider analysis and unsupervised classification of electroencephalography (EEG) data—an application area with currently very limited use of unsupervised approaches such as clustering. For example, EEG data analysis in the rapidly developing area of brain-computer interface (BCI) (*27*, *28*) is usually performed by (semi-)supervised methods such as independent component analysis (ICA), linear discriminant analysis (LDA), support vector machines (SVMs), decision trees, or artificial neural networks (ANN) [for review, please see (*29*–*31*)]. The BCI technology enables people to communicate with their environment or to activate and control certain devices solely by using the brain's neural activity. Herewith, it opens a perspective for restoring motor ability or communication to severely disabled and paralyzed people as well as for ill individuals who experience significant physical limitations such as amyotrophic lateral sclerosis, locked-in syndrome, or other severe neuromuscular disorders (*32*–*34*). The benefits and accuracy of such a prosthetic system in the so-called noninvasive setting, however, heavily depend on correct recognition of patterns within the recorded signal as well as an unerring classification of the executed signals. This task is hampered by serious challenges: in contrast to the excellent (millisecond) temporal resolution of EEG, the spatial information of the neuronal activity is rather poor (resolution of centimeters).

The observable information, the so-called electroencephalogram, is the summed activity of about 10^6 to 10^8 neurons lying in the vicinity of the electrodes (*35*) being recorded and condensed by a comparatively tiny number of electrodes (normally between 64 and 128). This signal is also characterized by a poor signal-to-noise ratio and high uncertainty because scull, skin, and hair are damping and skewing the electromagnetic waves (*36*). Furthermore, the measured signal reflects both the intrinsic neuronal activity within the cerebral cortex as well as the nerve impulses received from subcortical structures and the sense receptors. Such a signal is by its nature nonstationary and nonlinear, which makes the analysis and classification of the underlying signal patterns very difficult. Systematically missing data and unresolved scales may result in a problem of nonstationarity (*37*, *38*), which may lead to biased results when using the most common state-of-the-art classifiers in BCI research (ICA, LDA, SVMs, or ANN) that are all based on some form of intrinsic stationarity assumptions (*38*, *39*).

Although a wide range of these classifiers have been successfully applied to several important problems such as the noninvasive identification of epileptic patterns, almost all of the classifications of (imaginary) left/right hand or foot movements, or classification of emotions (*29*–*31*), given the inter- and intrapersonal variations in EEG as well as the poor signal-to-noise ratio—require very long data sets for the initial training to obtain satisfactory performance. In addition, from the mathematical/computational perspective, they are based on a priori assumptions concerning the distance metric (mostly chosen to be Euclidean) and/or the probabilistic assumptions (for example, Gaussianity and independence). In practical applications, however, there is no guarantee that these necessary assumptions can be fulfilled a priori. For this purpose, as a practical illustration of the clustering methodology is introduced in this article, we are presenting a way of using nonparametric clustering approaches to enable classification of high-dimensional experimental EEG data without initial training or a priori probabilistic assumptions on the nature of the data. As will be demonstrated below on a particular set of EEG data, this procedure leads to a very accurate unsupervised classification of very short nonstationary and noisy EEG signals.

As a data source, we take the EEG Motor Movement/Imagery Dataset from Physionet.org (*40*), which was created using the BCI2000 instrumentation system (*28*) (www.bci2000.org). This data set consists of more than 1500 1- and 2-min EEG recordings obtained from 109 volunteers. Subjects performed different motor/imagery tasks while 64-channel EEG activity was recorded with electrodes positioned according to the international 10-10 system. Each subject performed—besides other experimental runs—the two 1-min baseline runs (one with eyes opened, one with eyes closed) when the basic activity of the brain is being measured. In the following, we will only focus on these baseline measurements. These measurements for opened and closed eyes are similar to such an extent that standard unsupervised methods (as will be seen later) are not able to distinguish between them.

The experiment is undertaken as follows: the subject (without any external disturbance) is sitting for 1 min with closed eyes and for another minute with opened eyes. Each of the two investigated experiments (baseline "eyes opened" and baseline "eyes closed") has 64 dimensions (due to the 64 electrodes), with U = 9760 measurements per minute (which makes a demand interval of about 160 Hz). Thus, for every experimental run, we have to deal with a matrix in $R^{64\times U}$. To ensure the comparability of different EEG data sets and applicability of Euclidean distance measure, a series of preprocessing steps (including differencing, embedding, and dimension reduction of the embedded signal) has been performed (please see the section "Preprocessing" in the Supplementary Text for the explicit description of the data preprocessing protocol). For a distance measure g to be deployed in the clustering, we choose a measure associated with the principal components analysis (PCA), that is, the Euclidean distance

$$g(x(u), \Theta_i) = \|x(u) - \Theta_i \Theta_i^T x(u)\|_2^2$$

between the n-dimensional data points $x(u)$ and their orthogonal projections on the d-dimensional ($d < n$) locally linear manifold $\Theta_i \in \mathbf{R}^{n\times d}$ correspondent to the cluster i. For more details on PCA-induced clustering, please refer to (*16*, *23*). Next, we choose an appropriate initial information/assumptions that can be imposed on the clustering of the EEG data. We choose a very mild assumption that the underlying essential dynamics (captured as the temporal change of the low-dimensional manifold $\Theta(u) = \sum_{i=1}^{\mathbf{K}} \gamma_i(u)\Theta_i$) is a persistent and slowly varying process in time. Then, the a priori graph G is a linear time graph, and a matrix of weights W can be chosen, for example, as $W_{u,v}^{\sigma} = \exp[\sigma(u-v)^2]$ (with $\sigma < 0$, please see Fig. 1A).

Outcome

We tested the methodology for this particular choice of distance measure g and weight matrix W with the first 20 subjects of the original data set. We achieved an exact and unsupervised classification of opened and closed eyes measurements for all of the cases with the measurement fragments that were down to 7 seconds short. We will show exemplarily results for the subject Nr. 1. Figure 1B demonstrates the mAIC curves for models with $\mathbf{K}$ = 1, 2, 3 clusters as a function of the regularization constant ϵ^2. The overall minimum is attained at the position $\mathbf{K}$ = 2, $\epsilon^2 \geq 10^5$. As can be seen from Fig. 1B, from the viewpoint of information theory, the optimal solution of clustering problems obtained with standard unregularized clustering methods (such as k-means and hierarchical clustering algorithms) is attained for $\mathbf{K}$ = 1, and allows no distinction between the two states (that is, between opened and closed eyes), and is inferior in terms of information contents to the solution of the regularized problem (Eq. 4) for a given set of data. By introducing regularization, the overall minimum is attained with two clusters ($\mathbf{K}$ = 2), where one cluster only corresponds to the opened eyes and the second to the closed. Both experiments are correctly classified to their respective manifolds and can be visualized by plotting the cluster affiliation function γ for the optimal result (please see fig. S1). The two identified attractive manifolds—each of which is characteristic for one experiment—can be visualized by plotting the first three dimensions (out of the 300 most significant dimensions) that were detected during the data reprocessing via PCA. Figure 2 gives an impression of the dynamics of the two systems in phase space. Both dynamic systems are essentially nonlinear oscillators. Although they behave similarly, one can see that the orientation of the planes in which the oscillations take place is different. These principal attractor manifolds are approximated and distinguished via linear projectors Θ_i deployed in our PCA-based regularized clustering procedure. However, because the dynamics are geared into each other, standard algorithms are incapable of correctly solving this clustering problem. This result demonstrates that deploying the manifold-based clustering combined with a priori persistency assumption for the underlying dynamics allows us to use the respective structure of the manifold as a classifier to determine whether the unlabeled short measurement belongs to a subject with opened or closed eyes. New data can be projected on the manifolds and (by means of proximity) assigned to one of them. Standard PCA clustering [with $g(x(u), \Theta_i) = ||x(u) - \Theta_i \Theta_i' x(u)||_2$ but without the graph-induced regularization] was, in contrast, not able to detect the two manifolds and proposed a common basis manifold for the two situations. That is, graph-induced regularization introduced in this paper appears to be essential for the correct unsupervised classification of these data sequences.

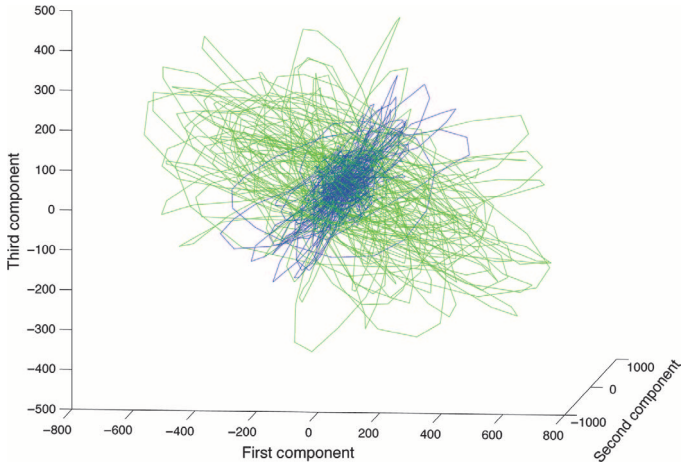

Fig. 2. Visualization of the two identified manifolds.

Revealing the essential spatiotemporal dynamics

With the help of the identified manifolds and their affiliated eigenvectors (constituting the columns of cluster projector matrices Θ_i), essential components of the underlying dynamical system can now be extracted from the available short, nonstationary, and noisy EEG time series. For this purpose, the experimental data are projected on the identified linear attractor manifolds Θ_i. Spectral analysis for the embedded projections of original EEG data on the dominant manifold dimensions (that is, on different columns of Θ_i as resulting from the regularized clustering) reveals the well-known α, β, γ, and μ waves of the brain (please see fig. S2). In contrast to the standard procedures of obtaining these signals (that involve a very long measurement series and a careful selection of points on the head where these measurements are performed), in the context of the presented methodology, these brain waves can be obtained from the full original EEG with a very short (down to 7 seconds) measurement length. In the next step, we are going to examine the spatiotemporal dynamics of these brain wave patterns. For this purpose, the snapshots of eigenvectors are visualized over a schematic representation of the head (indicating the positions and numbers of electrodes according to the international 10-10 system). Because of the embedding during the preprocessing, this visualization of embedded eigenvectors

(representing the dominant manifold components) results in spatiotemporal animations of the essential dimensions of the underlying dynamics that can be extracted from the two identified clusters. A selection of these animations is provided as movies S1 to S8. A couple of snapshots from movies S1 (animating the most dominant attractor dimension for the experiment with opened eyes) and S2 (animating the most dominant attractor dimension for the experiment with closed eyes) are exemplarily presented in Fig. 3. The left column of Fig. 3 presents a series of snapshots taken from movie S1. These snapshots capture the dynamics in the most dominant manifold dimension (that is, the first column of the obtained Θ_i in the respective cluster) for the experiment with opened eyes. This dominant dynamics—a spatiotemporal oscillation—takes place in the anterior part of the brain: most evident frontally and propagating into the central (and even posterior) regions of the brain. These spatial characteristics (together with the observation of fig. S2) allow to conclude that the observed dominant pattern reflects a combination of rhythmical β activity [which is usually encountered over the frontal and central brain regions (*41*)] and γ waves [mainly observed in the visual cortex (*42*, *43*)]. Furthermore, the movie reveals that the main spatiotemporal dynamics for the EEG with opened eyes can be explained by a traveling wave oscillating between the frontal and the posterior regions of the brain. This dynamic is hidden in the very noisy EEG signal and can be uncovered by the presented cluster analysis methodology.

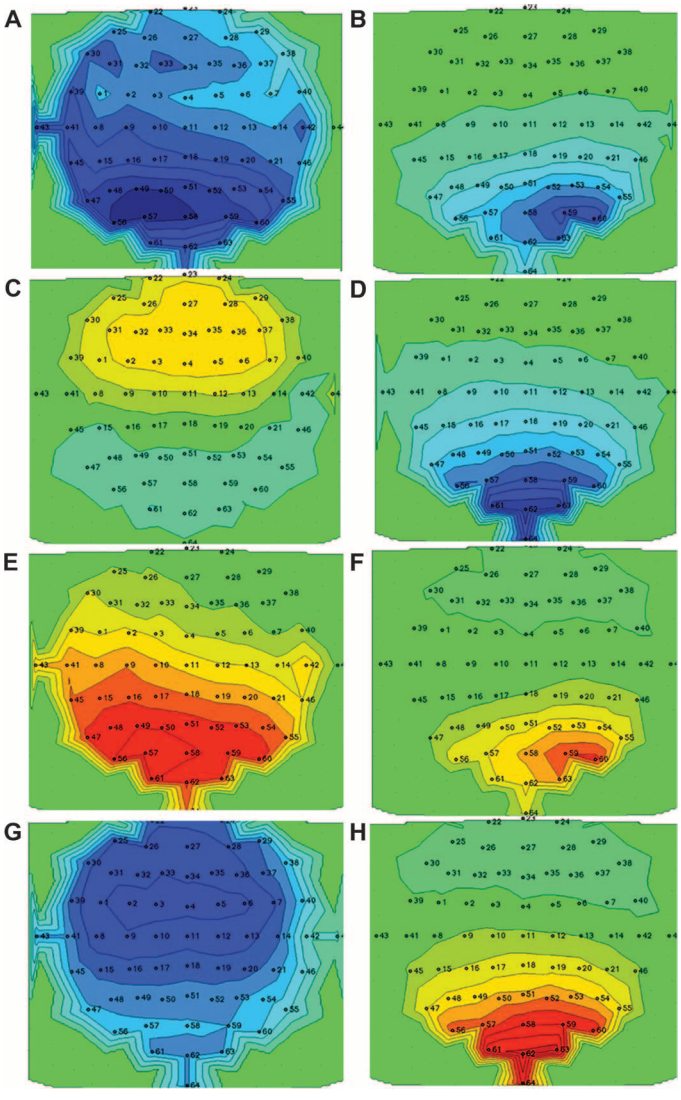

Fig. 3. Snapshots of the spatiotemporal dynamics of the most dominant eigenvectors. (**A**, **C**, **E**, and **G**) Several time instances of the extracted dominant wave pattern for the EEG with opened eyes. (**B**, **D**, **F**, and **H**) Snapshots for the dominant EEG pattern with closed eyes at the same time points. Red color stands for the positive component of the oscillation, and blue color for the negative component. Snapshots are taken in both experiments at time points t = 0.0 s (A and B), 0.018443 s (C and D), 0.043033 s (E and F), and 0.061475 s (G and H).

Entirely different dominant spatiotemporal dynamics are revealed in the situation with closed eyes (right column of Fig. 3): the oscillations are clearly located in the posterior half of the head only propagating to a minor degree into central areas (please see in addition movie S2 for the full animation). This pattern may be dedicated to the α rhythm, which is usually found over occipital, parietal, and posterior temporal regions of the brain (*41*, *44*). The animations of higher components (4, 7, and 10) for both experiments are provided by additional movies S3 to S8.

Concluding discussion

We have presented a methodology for imposing a priori graph/network information on clustering algorithms. The theoretical justification of this approach was presented, based on the derivation of a generalized graph/network-related regularization strategy, proving that it allows us to find the approximate solutions of various heterogeneous data analysis problems through upper-bound minimization and obtaining a particular analytic solution (Eq. 6). This thereby resolves a main computational bottleneck of a large family of time series clustering algorithms. The introduced approach is neither in competition with the various existing clustering methods nor is it just "yet another clustering method." As demonstrated above, presented methodology can be implemented as a very straightforward modification of the existent clustering algorithms for solving unsupervised and nonparametric classification problems.

When applying this approach to different EEG data sets, it was shown that combining the new methodology with Taken's theorem, data embedding, and PCA-related clustering, one can achieve the exact unsupervised classification of very short unlabeled EEG sequences. Standard clustering methods without regularization (for example, *k*-means, spectral clustering, and hierarchical clustering) were not able to differentiate between the two situations and suggested a common model/cluster, thus excluding the possibility for the unsupervised classification of the analyzed data. The introduced methodology operated as an exact classifier and detected two distinct clusters as the best model—each of which associated with only one of the two states. Furthermore, it was exemplified how the obtained results can be used to extract the dominant spatiotemporal dynamic patterns which are otherwise hidden in very noisy nonstationary signals. We also investigated the possibility of using the identified manifolds from one subject's EEG to classify the signal in the EEG of another subject. However, results differ significantly across probands, implying that the analysis must be uniquely performed separately for each individual.

The presented general framework for graph-induced regularization of clustering problems is expected to become helpful in the areas where the already collected information can be represented as a graph and deployed to increase the quality of clustering results. For example, this may be done in the areas of bioinformatics (where the a priori information is

put together in the form of a so-called gene ontology graph) or in geoscience (where the prior notion about scaling cascades and self-similarity can be represented as directed graphs, giving a possibility to deploy them in cluster analysis of geophysical data). A brief overview and classification of the clustering methods that can potentially profit from this methodology are given within the section "Classification of the clustering algorithms" in the Supplementary Text.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/1/7/e1500163/DC1

Text

Movie S1. 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.

Movie S2. 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.

Movie S3. 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.

Movie S4. 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.

Movie S5. 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.

Movie S6. 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.

Movie S7. Visualization of the 10th 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.

Movie S8. Visualization of the 10th 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.

Fig. S1. Cluster affiliation function for the two identified manifolds.

Fig. S2. Spectrograms of the EEG data.

References (*45*–*84*)

REFERENCES AND NOTES

1. A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: A review. *ACM Comput. Surv.* **31**, 264–323 (1999).
2. M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques. *J. Intell. Inf. Syst.* **17**, 107–145 (2001).
3. R. Xu, D. Wunsch II, Survey of clustering algorithms. *IEEE Trans. Neural. Netw.* **16**, 645–678 (2005).
4. B. Andreopoulos, A. An, X. Wang, M. Schroeder, A roadmap of clustering algorithms: Finding a match for a biomedical application. *Brief. Bioinform.* **10**, 297–314 (2009).
5. T. Parsons, Persistent earthquake clusters and gaps from slip on irregular faults. *Nat. Geosci.* **1**, 59–63 (2008).
6. A. Skelton, M. Andrén, H. Kristmannsdóttir, G. Stockmann, C.-M. Mörth, Á. Sveinbjörnsdóttir, S. Jónsson, E. Sturkell, H. Rakel Guðrúnardóttir, H. Hjartarson, H. Siegmund, I. Kockum, Changes in groundwater chemistry before two consecutive earthquakes in Iceland. *Nat. Geosci.* **7**, 752–756 (2014).
7. J. D. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384 (1989).
8. G. Leibon, S. Pauls, D. Rockmore, R. Savell, Topological structures in the equities market network. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 20589–20594 (2008).
9. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863–14868 (1998).
10. W. Huang, X. Cao, F. H. Biase, P. Yu, S. Zhong, Time-variant clustering model for understanding cell fate decisions. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E4797–E4806 (2014).
11. S. Galbraith, J. A. Daniel, B. Vissel, A study of clustered data and approaches to its analysis. *J. Neurosci.* **30**, 10601–10608 (2010).
12. D. Allen, G. Goldstein, *Cluster Analysis in Neuropsychological Research Recent Applications* (Springer, New York, 2013).
13. J. Hartigan, *Clustering Algorithms, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics* (Wiley, New York, 1975).
14. F. Höppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition* (John Wiley & Sons, New York, 1999).
15. J. C. Bezdek, R. Ehrlich, W. Full, FCM: The fuzzy *c*-means clustering algorithm. *Comput. Geosci.* **10**, 191–203 (1984).
16. P. Metzner, L. Putzig, I. Horenko, Analysis of persistent nonstationary time series and applications. *Comm. App. Math. Comp. Sci.* **7**, 175–229 (2012).
17. A. N. Tihonov, On the solution of ill-posed problems and the method of regularization. *Dokl. Akad. Nauk SSSR* **151**, 501–504 (1963).
18. A. Tikhonov, V. Arsenin, *Solutions of Ill-Posed Problems* (Winston, New York, 1977).
19. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
20. E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen, *LAPACK Users' Guide* (Society for Industrial and Applied Mathematics, Philadelphia, PA, ed. 3, 1999).
21. L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, R. C. Whaley, ScaLAPACK Users' Guide (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997).
22. M. Castillo, E. Chan, F. D. Igual, R. Mayo, E. S. Quintana-Orti, G. Quintana-Orti, R. van de Geijn, and F. G. Van Zee, "FLAME working note 31: Making parallel programming synonymous with programming for linear algebra libraries" (Technical Report TR-08-20, The University of Texas at Austin, Austin, TX, 2009).
23. I. Horenko, Finite element approach to clustering of multidimensional time series. *SIAM J. Sci. Comput.* **32**, 62–83 (2010).
24. I. Horenko, On the identification of nonstationary factor models and their application to atmospheric data analysis. *J. Atmos. Sci.* **67**, 1559–1574 (2010).
25. H. Akaike, A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
26. A. J. Majda, X. Wang, *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows* (Cambridge Univ. Press, Cambridge, 2006).
27. J. Wolpaw, E. W. Wolpaw, *Brain-Computer Interfaces: Principles and Practice* (Oxford Univ. Press, New York, 2012).
28. G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, J. Wolpaw, BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.* **51**, 1034–1043 (2004).
29. U. R. Acharya, S. V. Sree, G. Swapna, R. J. Martis, J. S. Suri, Automated EEG analysis of epilepsy: A review. *Knowl. Based Syst.* **45**, 147–165 (2013).
30. F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, B. Arnaldi, A review of classification algorithms for EEG-based brain–computer interfaces. *J. Neural Eng.* **4**, R1–R13 (2007).
31. S. M. Khorshidtalab A, 4th International Conference on Mechatronics (ICOM), IEEE, Kuala Lumpur, Malaysia, 17 to 19 May 2011, pp. 1–7.
32. N. Birbaumer, A. Murguialday, L. Cohen, Brain-computer interface in paralysis. *Curr. Opin. Neurol.* **21**, 634–638 (2008).
33. R. Rupp, Challenges in clinical applications of brain computer interfaces in individuals with spinal cord injury. *Front. Neuroeng.* **7**, PMC4174119 (2014).
34. J. Höhne, E. Holz, P. Staiger-Sälzer, K.-R. Müller, A. Kübler, M. Tangermann, Motor imagery for severely motor-impaired patients: Evidence for brain-computer interfacing as superior control solution. *PLOS One* **9**, e104854 (2014).
35. J. W. Sleigh, D. A. Steyn-Ross, M. L. Steyn-Ross, C. Grant, G. Ludbrook, Cortical entropy changes with general anaesthesia: Theory and experiment. *Physiol. Meas.* **25**, 921–934 (2004).
36. P. Herman, G. Prasad, T. McGinnity, 27th Annual International Conference of the Engineering in Medicine and Biology Society, Shanghai, China, 1 to 4 September 2005, pp. 5354–5357.
37. S. Gerber, I. Horenko, On inference of causality for discrete state models in a multiscale context. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14651–14656 (2014).
38. J. de Wiljes, L. Putzig, I. Horenko, Discrete nonhomogeneous and nonstationary logistic and Markov regression models for spatiotemporal data with unresolved external influences. *Comm. App. Math. Comp. Sci.* **9**, 1–46 (2014).
39. O. Kaiser, I. Horenko, On inference of statistical regression models for extreme events based on incomplete observation data. *Comm. App. Math. Comp. Sci.* **9**, 143–174 (2014).
40. J. J. Goldberger, S. Challapalli, M. Waligora, A. H. Kadish, D. A. Johnson, M. W. Ahmed, S. Inbar, Uncertainty principle of signal-averaged electrocardiography. *Circulation* **101**, 2909–2915 (2000).
41. E. Neidermeyer, The normal EEG of the waking adult, in *Electroencephalography: Basic Principles, Clinical Applications and Related Fields*, E. Niedermeyer, F. H. Lopes da Silva, Eds. (Williams and Wilkins, Baltimore, 1999), pp. 149–173.
42. J. Swettenham, S. Muthukumaraswamy, K. Singh, Spectral properties of induced and evoked gamma oscillations in human early visual cortex to moving and stationary stimuli. *J. Neurophysiol.* **102**, 1241–1253 (2009).
43. S. Muthukumaraswamy, K. Singh, Spatiotemporal frequency tuning of BOLD and gamma band MEG responses compared in primary visual cortex. *Neuroimage* **40**, 1552–1560 (2008).

44. E. Adrian, B. Matthews, The Berger rhythm: Potential changes from the occipital lobes in man. *Brain* **57**, 355 (1934).
45. I. Horenko, C. Schütte, On metastable conformational analysis of nonequilibrium biomolecular time series. *SIAM Mult. Mod. Sim.* **8**, 701 (2010).
46. I. Horenko, On clustering of non-stationary meteorological time series. *Dyn. Atmos. Oceans* **49**, 164–187 (2010).
47. R. Coifman, S. Lafon, Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30 (2006).
48. D. Zhou, B. Schölkopf, T. Hofmann, *Advances in Neural Information Processing Systems (NIPS)*, L. Saul, Y. Weiss, L. Bottou, Eds. (MIT Press, Cambridge, 2005), vol. 17, pp. 1633–1640.
49. D. Zhou, J. Huang, B. Schölkopf, Proceedings of the 22nd International Conference on Machine Learning (ICML), L. D. Raedt, S. Wrobel, Eds. (ACM Press, New York, 2005), pp. 1041–1048.
50. H. Whitney, *The Collected Papers of Hassler Whitney. Volume I and II* (Birkhuser Verlag, Basel, 1992).
51. F. Takens, Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence, Lecture Notes in Mathematics* (Springer-Verlag, Berlin, 1980), vol. 898, pp. 366–381.
52. D. Broomhead, G. King, *Nonlinear Phenomena and Chaos*, S. Sarkar, Ed. (Adam Hilger, Bristol, 1986), pp. 113–144.
53. D. Broomhead, G. P. King, Extracting qualitative dynamics from experimental data. *Physica D* **20**, 217–236 (1986).
54. M. Ghil, R. Vautard, Interdecadal oscillations and the warming trend in global temperature time series. *Nature* **350**, 324–327 (1991).
55. M. Ghil, M. R. Allen, M. D. Dettinger, K. Ide, D. Kondrashov, M. E. Mann, A. W. Robertson, A. Saunders, Y. Tian, F. Varadi, P. Yiou, Advanced spectral methods for climatic time series. *Rev. Geophys.* **40**, 3.1–3.41 (2002).
56. B. Schölkopf, A. Smola, E. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319 (1998).
57. I. Horenko, On simultaneous data-based dimension reduction and hidden phase identification. *J. Atmos. Sci.* **65**, 1941–1954 (2008).
58. D. L. Donoho, C. Grimes, Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5591–5596 (2003).
59. P. Mahalanobis, On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **2**, 49–55 (1936).
60. J. B. MacQueen, Proceedings of the Fifth Symposium on Mathematical Statistics and Probability (University of California Press, Berkeley, CA, 1967).
61. M. A. Hartigan, J. A. Wong, Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 100–108 (1979).
62. L. Kaufman, P. Rousseeuw, Clustering by means of medoids, in *Statistical Data Analysis Based on the L1Norm, Reports of the Faculty of Mathematics and Informatics. Delft University of Technology* (Elsevier, Amsterdam, 1987), pp. 405–406.
63. P. S. Bradley, O. L. Mangasarian, W. N. Street, *Advances in Neural Information Processing Systems 9* (MIT Press, Cambridge, MA, 1997), pp. 368–374.
64. D. Arthur, S. Vassilvitskii, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '07* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 2007), pp. 1027–1035.
65. S. Johnson, Hierarchical clustering schemes. *Psychometrika* **2**, 241–254 (1967).
66. T. Zhang, R. Ramakrishnan, M. Livny, BIRCH: A new data clustering algorithm and its applications. *Data Min. Knowl. Discov.* **1**, 141–182 (1997).
67. S. Guha, R. Rastogi, K. Shim, Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, 1998.
68. G. Karypis, Eui-Hong, S. Han, V. Kumar, Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Comput.* **32**, 68–75 (1999).
69. L. Baum, An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities* **3**, 1–8 (1972).
70. I. Horenko, E. Dittmer, C. Schütte, Reduced stochastic models for complex molecular systems. *SIAM Comp. Vis. Sci.* **9**, 89–102 (2005).
71. J. Banfield, A. Raftery, Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803–821 (1993).
72. A. Gelman, J. Carlin, H. Stern, D. Rubin, *Bayesian Data Analysis* (Chapman and Hall, London, 2003).
73. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **39**, 1–38 (1977).
74. T. Gasser, H. G. Muller, Estimating regression functions and their derivatives by the kernel method. *Scand. J. Stat.* **11**, 171–185 (1984).
75. C. Loader, *Local Regression and Likelihood (Statistics and Computing)* (Springer, New York, 1999).
76. T. Kohonen, Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
77. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, 1996, pp. 226–231.
78. J. Sander, M. Ester, H.-P. Kriegel, X. Xu, Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Min. Knowl. Discov.* **2**, 169–194 (1998).
79. M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander, Proceedings of the ACM SIGMOD International Conference on Management of Data, Philadelphia, PA, 1999.
80. R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, Automatic subspace clustering of high dimensional data. *Data Min. Knowl. Discov.* **11**, 5–33 (2005).
81. A. Rodriguez, A. Laio, Machine learning. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
82. U. von Luxburg, A tutorial on spectral clustering, *Stat. Comput.* **17**, 395–416 (2007); http://dx.doi.org/10.1007/s11222-007-9033-z.
83. W. Wang, J. Yang, R. Muntz, Proceedings of 23rd Very Large Data Bases Conference, Athens, Greece, 1997.
84. C. Sheikholeslami, S. Chatterjee, A. Zhang, Proceedings of 24th Very Large Databases Conference, New York, NY, 1998, pp. 428–439.

Acknowledgments: Both authors contributed equally to this study. **Funding:** The work of S.G. and I.H. is partly funded by the Swiss National Research Foundation (grant 200021 152979), Swiss Platform for Advanced Scientific Computing, and German Research Foundation ("Mercator Fellowship" of I.H. in the CRC 1114 "Scaling Cascades in Complex Systems"). **Competing interests:** The authors declare that they have no competing interests.

Submitted 5 February 2015
Accepted 25 June 2015
Published 7 August 2015
10.1126/sciadv.1500163

Citation: S. Gerber, I. Horenko, Improving clustering by imposing network information. *Sci. Adv.* **1**, e1500163 (2015).

ScienceAdvances

www.advances.sciencemag.org/cgi/content/full/1/7/e1500163/DC1

Supplementary Materials for

Improving clustering by imposing network information

Susanne Gerber and Illia Horenko

Published 7 August 2015, *Sci. Adv.* **1**, e1500163 (2015)
DOI: 10.1126/sciadv.1500163

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.advances.sciencemag.org/cgi/content/full/1/7/e1500163/DC1)

Movie S1 (.avi format). 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.
Movie S2 (.avi format). 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.
Movie S3 (.avi format). 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.
Movie S4 (.avi format). 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.
Movie S5 (.avi format). 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.
Movie S6 (.avi format). 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.

Movie S7 (.avi format). Visualization of the 10th 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.
Movie S8 (.avi format). Visualization of the 10th 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.

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

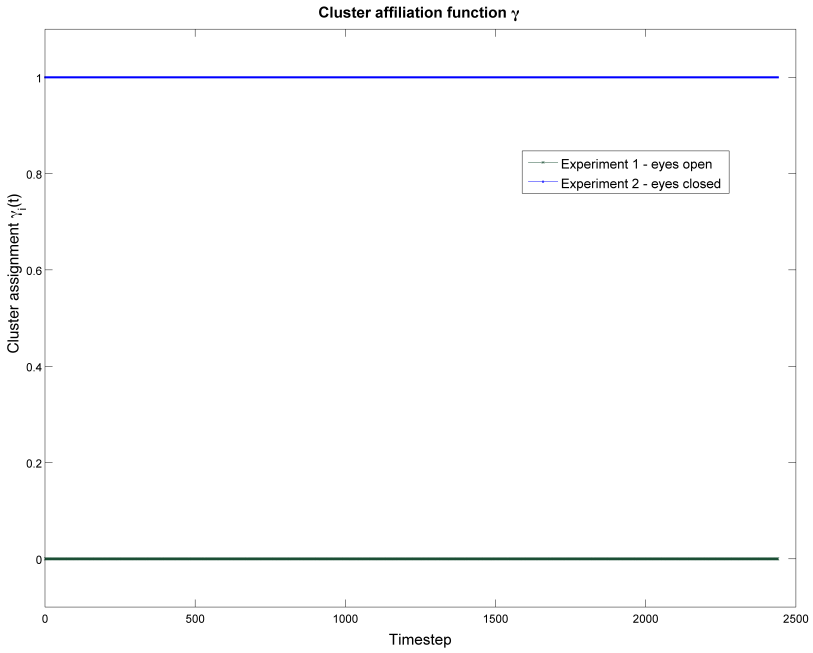

Figure S 1: Cluster affiliation function for the two identified manifolds

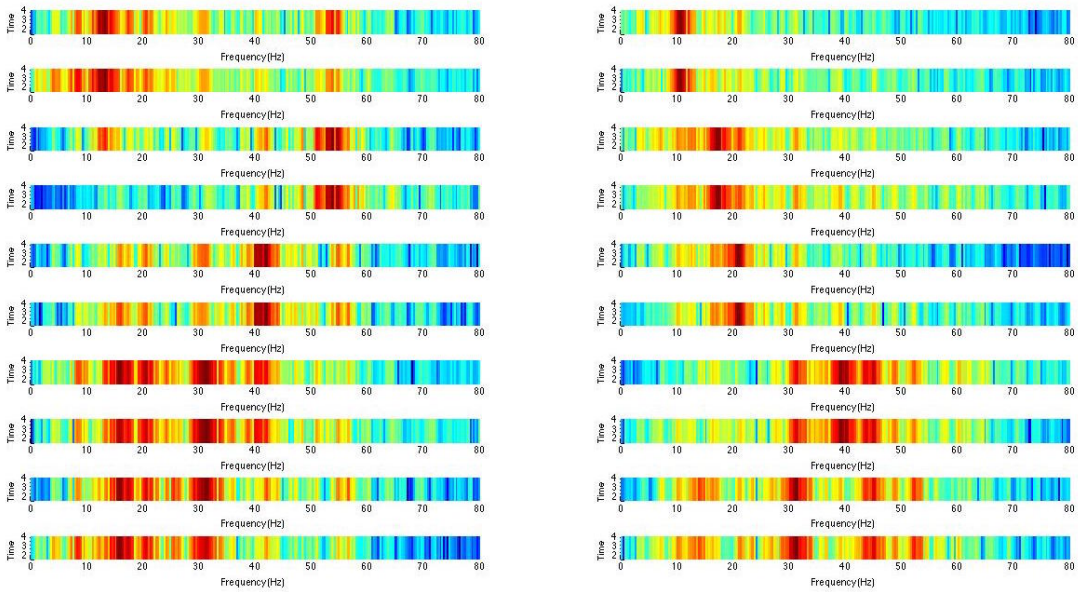

Figure S 2: Spectrograms of the EEG data. The left column shows the spectrogram of the projected experiment "eyes open" and the right column the spectrogram of projected experiment "eyes closed". Each row is computed from the projection of the EEG data on one of the first 10 dominant dimensions of the 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.