Mining Interaction Patterns among Brain Regions by Clustering

Claudia Plant, Andrew Zherdin, Christian Sorg, Anke Meyer-Baese, and Afra M. Wohlschläger

Abstract—Functional magnetic resonance imaging (fMRI) provides the potential to study brain function in a non-invasive way. Massive in volume and complex in terms of the information content, fMRI data requires effective, and efficient data mining techniques. Recent results from neuroscience suggest a modular organization of the brain. To understand the complex interaction patterns among brain regions we propose a novel clustering technique. We model each subject as multivariate time series, where the single dimensions represent the fMRI signal at different anatomical regions. In contrast to previous approaches, we base our cluster notion on the interactions between the univariate time series within a data object. Our objective is to assign objects exhibiting a similar intrinsic interaction pattern to a common cluster. To formalize this idea, we define a cluster by a set of mathematical models describing the cluster-specific interaction patterns. Based on this novel cluster notion, we propose interaction K-means (IKM), an efficient algorithm for partitioning clustering. An extensive experimental evaluation on benchmark data demonstrates the effectiveness and efficiency of our approach. The results on two real fMRI studies demonstrate the potential of IKM to contribute to a better understanding of normal brain function and the alternations characteristic for psychiatric disorders.

Index Terms—Clustering, multivariate time series, interaction patterns

1 Introduction

Human brain activity is very complex and far from being fully understood. Many psychiatric disorders like Schizophrenia and Somatoform Pain Disorder can so far neither be identified by biomarkers, nor by physiological or histological abnormalities of the brain. Aberrant brain activity often is the only resource to understand psychiatric disorders. Functional magnetic resonance imaging (fMRI) opens up the opportunity to study human brain function in a non-invasive way. The basic signal of fMRI relies on the blood-oxygen-level-dependent (BOLD) effect, which allows indirectly imaging brain activity by changes in the blood flow related to the energy consumption of brain cells. In a typical fMRI experiment, the subject performs some cognitive task while in the scanner. Recently, resting-state fMRI has attracted considerable attention in the neuroscience community [1]. Surprisingly, only about 5% of the energy consumption of the human brain can be explained by the task-related activity. Many essential brain functions, e.g. long-term memory are largely happening during rest, most of them without consciousness of the subject and many of them are still not well understood. Therefore recent findings support the potential of resting-state fMRI to explore the brain function in healthy subjects and reveal alternations characteristic for psychiatric disorders (e.g. [2]) . In resting state fMRI, subjects are instructed to just close their eyes and relax while in the scanner. fMRI data are time series of 3-dimensional volume images of the brain. The data is traditionally analyzed within a mass-univariate framework essentially relying on classical inferential statistics, e.g. contained in the software package SPM [3]. A typical statistical analysis involves comparing groups of subjects or different experimental conditions based on univariate statistical tests on the level of the single 3-d pixels called voxels. Data from fMRI experiments are massive in volume with more than hundred thousands of voxels and hundreds of time points. Since these data represent complex brain activity, also the information content can be expected to be highly complex. Only a small part of this information is accessible by univariate statistics. To make more of the potentially available information accessible, we need effective and efficient multivariate data mining methods.

Recent findings suggest a modular organization of the brain into different functional modules [4]. To obtain a better understanding of complex brain activity, it is essential to understand the complex interplay among brain regions during task and at rest. Inspired by this idea, we propose a novel technique for mining the different interaction patterns in healthy and diseased subjects by clustering. At the core of our method is a novel cluster notion: A cluster is defined as a set of subjects sharing a similar interaction pattern among their brain regions. After standard preprocessing including parcellation into anatomical regions, we model each subject as a data object which is represented by a multivariate time series. Each of the dimensions is a time series corresponding to the fMRI signal of a specific anatomical brain region. Our approach Interaction K-means

- *C. Plant and A. Meyer-Baese are with the Department of Scientific Computing, Florida State University, Tallahassee, FL 32306-4120 USA. E-mail: {cplant, ameyerbaese}@fsu.edu.*
- *A. Zherdin, C. Sorg, and A. M. Wohlschläger are with the University Hospital "Rechts der Isar" Technische Universität (TUM) Munich, 81675 Munich, Germany. E-mail: {andrew.zherdin, csorg, wohlschlaeger}@lrz.tum.de.*

Manuscript received 23 Mar. 2012; revised 5 Apr. 2013; accepted 7 Apr. 2013. Date of publication 18 Apr. 2013; date of current version 31 July 2014. Recommended for acceptance by M. Ester.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier 10.1109/TKDE.2013.61

(IKM) simultaneously clusters the data and discovers the relevant cluster-specific interaction patterns. The algorithm IKM is a general technique for clustering multivariate time series and not limited to fMRI data. Besides fMRI, multivariate time series are prevalent in many other applications. Increasing amounts of motion stream data are collected in multimedia applications [5]. Gesture sensing devices, such as a CyberGlove usually contain multiple sensors to capture human movements. Human motion stream data can also be extracted from video streams. In this application, it makes sense to regard each movement as a data object. A cluster analysis of motion stream data potentially identifies clusters with similar movements, usually performed by different persons.

Clustering time series has already reached high maturity with multiple books and book chapters [6], surveys [7], [8] and a huge volume of research papers, e.g., [9]–[16], to mention a few. Most of the mentioned techniques consider either the time series as whole as the data objects to be subjected to a cluster analysis or perform clustering on subsequences which allows defining more meaningful similarity measures in many applications. Defining a meaningful similarity measure is a non-trivial task and probably the most important challenge in clustering time series. In [17] Lin and Keogh illustrate some pitfalls associated with clustering based on subsequences with Euclidean distance and propose an alternative similarity measure based on characteristic recurrent motifs.

Considering the practical relevance of clustering multivariate time series, only disproportionally few papers address this issue [16], [18]–[20]. Many of the univariate methods mentioned so far can be straightforwardly extended to the multivariate case. However, by doing so, information is lost: Data which is inherently multivariate often contains interactions between the different time series. We demonstrate that this information can be very useful for clustering. In our examples, this aspect is intuitively reasonable: A motion is characterized by a specific pattern of dependencies among the recording sensors. A diseased person has a characteristic interaction pattern of brain regions which differs from the healthy controls. In [21] we first introduced this novel cluster notion. In this paper, we extend the basic idea to support nonlinear models and demonstrate its potential to discover interaction patterns among brain regions from fMRI data.

1.1 Contributions

The major contributions of this paper can be summarized as follows:

1) We introduce a novel cluster notion for clustering multivariate time series based on attribute interactions.
2) We propose Interaction K-means (IKM), a partitioning clustering algorithm suitable to detect clusters of objects with similar interaction patterns.
3) We demonstrate that the information on interaction patters provides valuable insights for interpretation.
4) Motivated by a real challenge from a neuroscience application, IKM outperforms state-of-the-art techniques for clustering multivariate time series on synthetic data as well as on benchmark data sets from different applications.
5) On fMRI data from studies on Somatoform Pain Disorder and Schizophrenia our algorithm detects very interesting and meaningful interaction patterns.

Notations. We consider a data set DS with n objects. Each object $O \in DS$ is a d-dimensional multivariate time series. Each dimension or attribute $a_i \in \mathcal{A}$ is a time series with m time points, i.e. $a_i = \{t_{1,i}, \ldots, t_{m,i}\}$. We also use $\vec{a}_i$ to denote the m time points of dimension a_i as a column vector. We use italics to denote sets, e.g. $\mathcal{A}$ denotes the set of attributes of DS and $\mathcal{O}$ a set of objects. Capital letters denote matrices composed by column vectors of dimensions. We further denote by m^* the overall number of time points considering one distinct dimension of some fixed set of objects. We consider a clustering $\mathcal{C}$ as a non-overlapping partitioning of DS into K clusters, i.e. $DS = \bigcup_{1 \le j \le K} C_j$ and drop the indices whenever non ambiguous.

The remainder of this paper is organized as follows. In the next section, we briefly survey related work. In Section 3 we introduce the interaction-based cluster notion. Section 4 extends the cluster model to support nonlinear interactions. In Section 5 we propose the corresponding clustering algorithm Interaction K-means (IKM). Section 7 contains an extensive experimental evaluation with diverse datasets and Section 8 shows the results on neuroscience data. Section 9 concludes the paper.

2 Related Work

Time series clustering is a dynamic field with applications in medicine, astronomy and economics. In spite of the prevalence of multivariate time series only a few algorithms ([16], [18], [19]) [20] have been especially designed for their clustering. However, most methods for clustering univariate time series can be applied to multivariate time series as well. Therefore we provide a brief survey on these techniques.

The most difficult task in clustering time series is to find an appropriate similarity measure. Many approaches rely on feature transformation and dimensionality reduction. Features derived by Discrete Wavelet Transform and the Discrete Fourier Transform [22], as well as obtained by Principle Component Analysis [23], [24] have been successfully applied for clustering. Alternative approaches to feature extraction include e.g. the method of multiresolution piecewise aggregate approximation presented in [13]. Recently, compression-based similarity measures have been proposed, e.g. [14]. Each time series is represented by a bit string indicating the intervals where the value of the time series is above the mean value of the time series. This representation is used to compute an approximate clustering. The bit level representations are compressed using standard compression algorithms in order to reduce the I/O cost and to speed-up the clustering task. A compression-based similarity measure is also proposed in [25] to compare long time series structure using co-compressibility as a dissimilarity measure. The authors report impressive results in many applications, but this technique requires certain statistical conditions from data.

Many approaches use Hidden Markov Models (HMM) for clustering time series, e.g. [15], [19], [26], to mention a few. The concept of HMM is particularly useful to model temporal correlations among the measurements. The approach [15] for example uses segmental semi-Markov models. First, the time series are modeled by k distinct segments and then a Viterbi-like algorithm is applied to compute the similarity measure. In the experimental section, we compare to the Sequence Cluster Refinement Algorithm (SCRA) [18]. Like IKM, this approach is an iterative partitioning K-means-style clustering method but it uses HMMs to represent the cluster centers. In the assignment step, each time series object O is assigned to that cluster with the HMM model that most likely generated the sequence of O. In the update step, the HMM of each cluster is recomputed maximizing the probability of producing all time series currently assigned to the cluster. In contrast to HMMs, our models capture dependencies not between time points but between the dimensions. Our models are applicable to groups of objects and not to single objects. To describe the interaction patterns within the time series of an object, we apply linear models as recently proposed in [27], [28] for the purpose of efficient compression and classification of time series, respectively.

Some approaches especially focus on clustering multivariate time series. The authors of [20] suggest merging all dimensions of multivariate time series into a long univariate time and supplying it to the Autoclass algorithm [29]. This technique is not applicable if the number of time points varies among the objects, a case frequently occurring in fMRI data. In [18], the authors apply Independent Component Analysis to transform multivariate time series data into statistically independent components, and then propose a clustering algorithm called ICACLUS to group the data according to the ICs found. This approaches is based on the assumption that the observed multivariate time series are mixtures of statistically independent sources which need to be de-mixed in order to find the clusters. In the experimental section, we compare to ICACLUS, since Independent Component Analysis and related blind signal source separation techniques have been successfully applied for analyzing neuroimaging data, see e.g [30]. We further compare our method to structure-based statistical features clustering (SF) [16] which has been especially designed for clustering multivariate time series. SF represents each attribute of the multivariate time series by a fixed-length vector whose components are statistical features of the time series which capture the global structure. Statistical features include e.g. hurst, kurtosis and nonlinearity. These descriptive vectors, one for each component of the multivariate time series, are concatenated, before being clustered using a standard clustering algorithm such as K-means. Extensive experiments demonstrate that SF is suitable to identify activity patterns from motion stream data with high accuracy.

3 Interaction-Based Cluster Notion

In this section, we elaborate our cluster notion based on characteristic interaction patterns. We want to find clusters of objects which are represented by multivariate time series sharing a common cluster-specific interaction pattern among the dimensions. For an example consider Fig. 1 displaying a data object represented by a multivariate time series. The dimensions of this object exhibit a simple interaction pattern: The time series of dimension dim_{12} can be expressed by a linear combination of some other dimensions: $dim_{12} := 2 \cdot dim_4 + dim_5 + dim_6$. Typically, not all dimensions of an object are interacting. For simplicity, only the dimensions involved in the interaction pattern are displayed Fig. 1. Note that this is an overly simple example, but more complex patterns are prevalent in real-world data, and to the best of our knowledge, have not been exploited for clustering so far. Often, dependencies among dimensions are rather regarded as non-necessary redundant information which should be removed before clustering, e.g. by the application of PCA for linear dependencies. In this work, we want to preserve the information on attribute dependencies and demonstrate that it can indeed be very valuable for clustering. The resulting clusters are composed of objects exhibiting similar dependencies among their attributes which can therefore be interpreted as cluster-specific interaction patterns.

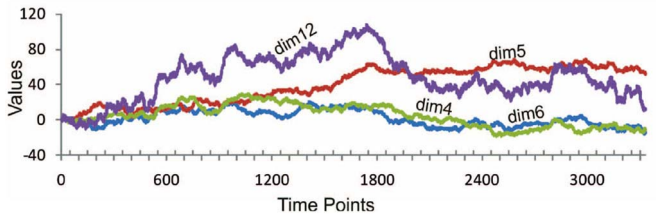

Fig. 1. Interaction pattern within a multivariate time series: The signal of dim_{12}, in our application representing a specific region of the brain, is provided by a linear combination of some other dimensions, here by dim_4 to dim_6, representing an interaction pattern among a set of brain regions.

To formally represent this idea, we define each cluster by a system of mathematical models. We first focus on linear models. Linear models have been successfully applied in a wide range of applications because they are interpretable and computationally efficient. In Section 4, we consider nonlinear dependencies, which can often be expressed by linearizable functions.

Definition 1 (Linear Model). *A linear model M_a for a dimension a and a set of objects $\mathcal{O}$ is provided by:*

$$A = V \cdot P + E, \text{ where}$$

- $A \in \mathbb{R}^{m^*}$ *results from horizontally concatenating the time series of dimension a of all objects in $\mathcal{O}$,*
- $V \in \mathbb{R}^{m^* \times |\mathcal{V}|}$ *contains the explanatory variables,*
- $P \in \mathbb{R}^{|\mathcal{V}|}$ *contains the parameters or coefficients of the model.*
- *and $E = \sqrt{||A - VP||^2} \in \mathbb{R}^{m^*}$ contains the error.*

We further denote sum of errors of the set of objects $\mathcal{O}$ with respect to M_a by $\mathcal{E}_{M_a} = \sum_{i=1}^{m^*} E_i$. We can also determine the error of an arbitrary object $P \notin \mathcal{O}$ as $\mathcal{E}_{P,M_a} = \sqrt{||A_P - V_P P||^2}$. A_P denotes the matrix obtained by projecting object P to dimension a. Analogously, V_P is the matrix containing the set of explanatory variables to model object P.

To obtain a single model for each dimension of a set of objects, we concatenate the time series of all objects. Note

that our method does not require that all time series of all objects are of equal length. To illustrate this definition, consider the model for dimension dim_{12} in Fig. 1 as an example : Assume we have 100 objects in a cluster. In this case, A is of size $333,300$, since the time series of each object consists of $3,333$ time points. The explanatory variables for dim_{12} are dim_4, dim_5 and dim_6. Therefore, matrix V is of size $333,300 \times 3$. Now, we are ready to define an interaction-based cluster.

Definition 2 (Interaction-based Cluster). *An interaction-based cluster C is defined by:*

- *A set of models $\mathcal{M}_C = \{M_{C,1}, \ldots, M_{C,d}\}$ representing the dependencies of each single dimension with respect to the other dimensions. We denote the model of dimension a_i of cluster C by $M_{C,i}$. For model $M_{C,i}$ we apply $\mathcal{V}_i \subseteqq \{\mathcal{A} \setminus a_i\}$ as set of explanatory variables in Definition 1. We will address the question how to find a suitable set of explanatory variables below.*
- *A set of data objects $\mathcal{O}_C$ associated to C.*
- *We denote the error of cluster C by $\mathcal{E}_C$ which is provided by $\mathcal{E}_C = \sum_{i=1}^{d} \mathcal{E}_{M_i}$.*
- *The error of object O with respect to cluster C, denoted by $\mathcal{E}_{O,C}$ is provided by $\mathcal{E}_{O,C} = \sum_{i=1}^{d} \mathcal{E}_{O,M_i}$.*

The aim of interaction-based clustering is to obtain a non-overlapping partitioning of *DS* into *K* clusters. Finally, each cluster should represent a specific interaction pattern which is characteristic for the assigned objects. More precisely, the optimization goal can be specified as follows:

Definition 3 (Clustering Objective Function). *The optimization goal for interaction-based clustering is to minimize the total sum of errors of all K clusters.*

$$\min \sum_{C \in \mathcal{C}} \frac{1}{m_C^*} \mathcal{E}_C.$$

To take into account that clusters consisting of objects with longer time series naturally tend to have a larger error, we normalize $\mathcal{E}_C$ by the total number of time points of the objects in cluster C , which we denote by m_C^*.

Model Finding. Before addressing the problem of how to find the clusters, we need to describe how the set of models $\mathcal{M}_C$ can be computed from the set of objects $\mathcal{O}_C$ which are associated to a cluster *C*. Since we focus on linear models, this involves solving d regression problems. Multiple least square regression can be applied to derive the models. However, a common problem is overfitting. The more dimensions are included into the model, the more variance is explained and thus the smaller is the error term. Models involving a large part or even all dimensions are not generalizable and hard to interpret. Therefore, the set of explanatory variables of each model needs to be carefully selected.

To determine the really relevant dimensions, we apply a greedy stepwise algorithm for model finding [31] in combination with the Bayesian Information Criterion (BIC) [32] as evaluation criterion. The greedy stepwise algorithm is an established technique for variable selection in regression problems. This algorithm starts with an empty set of relevant dimensions. In each step, either one dimension is added or removed, depending on which of these two actions is judged superior by the evaluation criterion. The algorithm terminates if none of the two actions leads to a further improvement. As evaluation criterion, we apply BIC which determines a balance between goodness-of-fit and complexity of the model and is defined by:

$$BIC(M_a) = -2 \cdot LL(a, M_a) + \log(m^*)(|\mathcal{V}| + 1).$$

The first term represents the goodness-of-fit, where $LL(a, M_a)$ denotes the log-likelihood of dimension a given the model. The second term punishes overly complex models.

4 Nonlinear Models

Interaction processes in nature are not limited to be linear. Therefore we extend our approach to also support nonlinear models.

Definition 4 (Nonlinear Model). *A nonlinear model M_a for a dimension a and a set of objects $\mathcal{O}$ is provided by:*

$$A = f(V, P) + E, \text{ where}$$

- *A, V, P specified as in Definition 1,*
- *$E = \sqrt{||A - f(V,P)||^2} \in \mathbb{R}^{m^*}$ contains the error;*
- *Π is the number of parameters for the function $f(\cdot, \cdot)$*
- *and $f(\cdot, \cdot)$ is an explanatory function: f:$\mathbb{R}^{m^* \times |\mathcal{V}|} \times \mathbb{R}^{\Pi} \rightarrow \mathbb{R}^{m^*}$.*

The definition is analog to Definition 1 but extends it for nonlinear cases. If $f(\cdot, \cdot)$ is a classical multiplication of the matrix and a vector and $\Pi = |V|$, Definition 2 coincides with Definition 1. Not every nonlinear model can be computed efficiently. We therefore consider only the sub-class of linearizable models. In linearizable models, each coefficient can be pre-computed and inserted as a new explanatory variable.

Definition 5 (Linearizable models). *A nonlinear model M_a for a dimension a and a set of objects $\mathcal{O}$ is provided by:*

$$A = \sum_{i=1}^{m'} f_i(v_i, p_i) + E, \text{ where}$$

- *A, V, P, E are specified analogously to Definition 1,*
- *m' is a number of summands,*
- *and $f_i(\cdot, \cdot)$ are a explanatory functions: f_i:$\mathbb{R}^{m^*} \times \mathbb{R} \rightarrow \mathbb{R}^{m^*}$*

A large class of nonlinear models can be specified by linearizable models. To give a simple example, consider the following:

$$A(t) = 3v_1^2(t) + 2\sin(v_2(t)) - log(v_3(t)).$$

We can rewrite this model as:

$$A(t) = 3\widetilde{v}_1(t) + 2\widetilde{v}_2(t) - \widetilde{v}_3(t)$$

with $\widetilde{v}_1(t) = v_1^2(t)$; $\widetilde{v}_2(t) = \sin(v_2(t))$; $\widetilde{v}_3(t) = log(v_3(t))$.

Due to the special properties of fMRI data, models with time-delay are especially useful. An event in a specific brain region measured by BOLD signal activity induces BOLD activity in another region by some time delay [33]. Particularly Granger causality between time series of BOLD activity is well suitable for modeling causal (i.e. time

algorithm IKM (data set DS, integer K): Clustering $\mathcal{C}$
  Clustering $bestClustering$;
//initialization
  for $init := 1 \ldots maxInit$ **do**
    $\mathcal{C} := randomInit(DS, K)$;
    for each $C \in \mathcal{C}$ **do**
      $\mathcal{M}_C := findModel(C)$;
    while not converged or iter $<$ **maxIter do**
//assignment
      for each $O \in DS$ **do**
        $O.cid = \min_{C \in \mathcal{C}} \mathcal{E}_{O,C}$
//update
      for each $C \in \mathcal{C}$ **do**
        $\mathcal{M}_C := findModel(C)$;
      if improvement of objective function
        $bestClustering := \mathcal{C}$;
    end while
  end for
return $bestClustering$;

Fig. 2. Algorithm interaction K-means.

delayed) influences of activity across remote brain areas, see [34]. To support integrating this aspect into our approach, we need need a linearizable model with time delay.

Definition 6 (Model with time delay). *A model with time delay M_a is nonlinear model with the explanatory function f provided by:*

$$f{:}\mathbb{R}^{m^* \times |\mathcal{V}|} \times \mathbb{R}^{|\mathcal{V}|} \rightarrow \mathbb{R}^{m^* - |\mathcal{O}|}$$

$$f(V, P){:} = \mathsf{TimeShift}_1(V) \cdot P \text{ where}$$

- V, P *are specified analogously to Definition 1*
- *and* $\mathsf{TimeShift}_1(V(t)){:} = V(t-1)$ *i.e. we cut off the last time point of each object from matrix V. The length of output time series is $m^* - |\mathcal{O}|$, where m is the length of input time series (here $|\mathcal{O}|$ is the number of objects in the model).*

We can describe time-delay model for one object with a following equation:

$$\mathsf{TimeShift}_{-1}(A) = \mathsf{TimeShift}_1(V)P + E \text{ or}$$

$$A(t) = V(t-1)P + E(t) \text{ for } \forall t \in [2 \ldots m^*].$$

Our experiments in Section 8.3 demonstrate that models with time delay can improve the clustering accuracy on fMRI data. Depending on the application domain, other model classes like threshold autoregressive models for financial applications could also be integrated.

5 Interaction K-Means Clustering

In this section, we introduce the algorithm interaction-K-means (IKM) which minimizes the clustering objective function provided in Definition 3. Similar to classical K-means [35], IKM is an iterative algorithm which efficiently converges towards a local minimum of the optimization space.

Algorithm IKM. Analogously to K-means, the first step of IKM is the *initialization*. As a common strategy for K-means, we propose to run IKM several times with different random initializations and keep the best overall result. For initialization, we randomly partition DS into K clusters. For IKM it is favorable that the initial clusters are balanced in size to avoid overfitting. Therefore, we partition the data set into K equally sized random clusters and find a set of models for each cluster as described in the previous section. After initialization, IKM iteratively performs the following two steps until convergence: In the *assignment* step, each object O is assigned to the cluster w.r.t. which the error is minimal, i.e. $O.cid = \min_{C \in \mathcal{C}} \mathcal{E}_{O,C}$. It is easy to see that this minimizes the objective function in Definition 3. After assignment, in the *update* step, the models of all clusters are reformulated. Pseudocode of IKM is provided in Fig. 2.

As an iterative partitioning clustering algorithm, IKM follows a similar algorithmic paradigm as K-means. However, note that there are significant differences: Our cluster notion requires a similarity measure, which is very different to LP metric distances. The similarity measure applied in IKM are the errors with respect to the set models of a cluster. This similarity measure is always evaluated between an object and a cluster, and not between two objects. In contrast to K-means or K-medoid algorithms, we can not state that a data object is the representative of a cluster. The cluster representative in IKM is a set of models describing a characteristic pattern of interaction among the dimensions.

Convergence. IKM converges as soon as no object changes its cluster assignment during two consecutive iterations. Usually, a fast convergence can be observed (less than 50 iterations on our experimental data), but there are some rare cases in which IKM does not converge. Analogously to standard K-means, it can be straightforwardly proven that the assignment and the update step strictly decrease the objective function provided in Definition 3. However, due to the greedy stepwise algorithm applied for model finding, see Section 3, the strictly monotonic property is lost. In particular, in different iterations of IKM, BIC may select different numbers of explanatory variables to be included in the cluster models. We therefore propose to terminate after *maxIter* iterations. Our experiments demonstrate that this has no negative effects on the quality of the clustering result. Recently, following the approach of smoothed analysis Arthur *et al*. [36] have proven that K-means converges in polynomial time on an arbitrary input data set subjected to random perturbations. This result theoretically supports our approach, since the fact that slightly different explanatory variable selection can happen in consecutive iterations is similar to minor perturbations of the data.

Runtime Complexity. As for ordinary K-means, the runtime of IKM scales linearly with the number of objects n, since the complexity of assignment step is linear

in n and usually only a few iterations are performed. Clearly, the update step is the most computationally expensive step, since model finding involves matrix inversion with complexity of $O(d^3)$ combined with the greedy stepwise algorithm with complexity of $O(d^2)$. Aggregative pre-computing inspired by [37] allows us to become virtually independent of the length of the time series. For model-finding we very frequently need to determine the model parameters P and the error E, which involves matrix multiplication of complexity $O(m^* d^2)$. Following [37], we pre-compute the matrix Θ^* with $\theta^*_{ij} = \theta^*_{ji} := \sum_{k=1}^{m^*} Z_{ki} \cdot Z_{kj}$, $Z := [V \mid A]$. With Θ^*, multiplication can be performed very efficiently:

$$V^T \cdot V = \left(\theta^*_{ij}\right), \quad \text{for } 1 \le i \le d-1, \quad 1 \le j \le d-1$$
$$V^T \cdot A = \left(\theta^*_{ij}\right), \quad \text{for } i = d, \quad 1 \le j \le d-1$$
$$P = (V^T V)(V^T A).$$

To determine the different models for the dimensions a cluster, we can always apply Θ^* and just need to apply another dimension as A. In contrast [37], we also need E. The computation of E can be supported by Θ^* by the following reduction:

$$\begin{aligned} E^2 &= ||A - VP|| = (A - VP)^T (A - VP) = \\ &= A^T A - A^T VP - (VP)^T A + (VP)^T (VP) = \\ &= A^T A - (V^T A)^T P - P^T (V^T A) + P^T (V^T V) P, \end{aligned}$$

with $(V^T A) \in \mathbb{R}^{(d-1)}$ and $(V^T V) \in \mathbb{R}^{(d-1)\times(d-1)}$, and $(A^T A) = \theta_{d,d} \in \mathbb{R}$, see Definitions 1, 2. Calculation of Θ^* needs $O(m^* \cdot d^2)$, but only once. Therefore, after determining the cluster model once, the costs for calculating the error are linar in the number of assigned objects. We calculate Θ^* as $\Theta^* = \sum_{i=1}^{|\mathcal{O}|} \Theta_i$. And Θ_i is precalculated for object i. $(\theta_i)_{kj} = \sum_{l=1}^{m} Z_{lk} \cdot Z_{lj}$. $|\mathcal{O}|$ is the number of objects in the cluster. The costs are amortized by each calculation of model coefficients and the error. Multiple runs of IKM increase this effect (see Section 7.3).

6 Interpretation of the Clustering Result

A major advantage of IKM is possibility to interpret the detected interaction patterns. To facilitate interpretation, we focus on a subset of the models which best differentiates among the clusters. For each pair of clusters, the best discriminating models are selected by leave-one-out validation using objects of the corresponding clusters. Fig. 3 displays the algorithm for interpretation in pseudocode. Considering a pair of clusters, we first generate the models of each individual cluster of from the training data. Then we compute the error of the test objects w.r.t. all models and sum up all errors. To obtain a ranking of the models regarding their ability to discriminate among the clusters, we consider errors w.r.t. the correct cluster of the test object with a positive sign (these errors should be small) and errors w.r.t the other cluster with a negative sign, respectively. Finally, we sort all models ascendingly according to the error. The top-ranked models best discriminate among the clusters.

User Feedback. The clustering result together with the information about which models best discriminate among clusters is a good basis for user interaction. Expert users can easily select the most relevant dimensions of the multivariate time series based on this information. Also, experts can easily verify their hypotheses on which dimensions, in the neuroscience application corresponding to anatomical brain regions are most relevant. After selecting the relevant regions, IKM can be run again. Our experiments in Section 8 demonstrate that this strategy can greatly improve the clustering result and thereby confirms hypotheses of the experts.

```
algorithm dimensionRanking
(Cluster C_i, Cluster C_j): ranking
  error_in_models := new ARRAY[d];
//leave-one-out-validation
  for each O ∈ O_{C_i} ∪ O_{C_j} do
    test := O
    O_{C_i} := O_{C_i} \ test; O_{C_j} := O_{C_j} \ test;
    findModel(C_i); findModel(C_j)
    for each cluster ∈ {C_i, C_j} do
      if O.cid = cluster.id then
        sign := 1;
      else
        sign := −1;
      end if
      for i := 1 ... d do
        error_in_models[d] += sign*
            cluster.models[d].
            calculateErrorFrom(O.
                getTimeSeries(d));
      end for
    end for
  end for
  sort(error_in_models);
return error_in_models;
```

Fig. 3. Algorithm for interpretation of the results.

7 Comparison to State-of-the-Art

7.1 Methodology

Selection of Comparison Methods. Existing approaches to clustering multivariate time series do not consider interaction information. Most approaches rather extract features from each dimension and cluster the resulting feature vector, see Section 2. To demonstrate that the information on attribute interaction is valuable for clustering multivariate time series, we compare IKM to two feature-based approaches. As a baseline, we consider classical K-Means clustering with Euclidean distance, which we term *Naive* in the following. The naive algorithm considers the concatenated dimensions of an object as a feature vector. In addition, we compare to *Statistical Feature Clustering (SF)* [16] which is a state-of-the-art feature-based approach to clustering multivariate time series, see Section 2. We compare to two further approaches addressing the challenge of clustering multi-variate time series in different non-feature-based ways: The *Sequence Cluster Refinement Algorithm (SCRA)* [18] uses Hidden

TABLE 1
Characteristics of Experimental Data Sets

name	type	domain	#objects	#cl.	#dim	length	reference
$DS1$	syn.	n.a.	600	6	13	3333	n.a.
$DS2$	syn.	n.a.	600	6	12	2048	n.a.
$DS3$	syn.	n.a.	120 - 640	6	12	2048	n.a.
$DS4$	syn.	n.a.	120	6	12 - 29	2048	n.a.
$DS5$	real	medical: EEG	20	2	64	256	[39]
$DS6$	real	motion stream	100	10	25	70	[40]
$DS7$	real	sp. recognition	640	9	12	7-29	[41]
$DS8$	syn.	n.a.	120	6	4 - 80	2048	n.a.
$DS9$	syn.	n.a.	120	6	12	100 - 4000	n.a.
$DS10$	syn.	n.a.	60 - 4200	6	12	2048	n.a.
$DS11$	real	medical: fMRI	26	2	90	216 or 325	[42]
$DS12$	real	medical: fMRI	26	2	5	180, 240, 300	[43]

Markov Models (HMMs) to represent clusters. The technique *ICACLUS* [19] relies on Independent Component Analysis (ICA), see Section 2.

Implementation, Parametrization and Data Sets. We implemented our method, the naive approach and ICACLUS in Java. We implemented SCRA in Matlab and obtained the R-code of SF from the authors. For ICACLUS, we used the FastIca algorithm [38] to perform ICA. For SCRA, we built the cluster models as super-models from dimension-wise classical HMMs[1] since the paper lacks an exact specification and we could not obtain the code from the authors.

All experiments were performed on a workstation equipped with a 2.66GHz CPU and 8GB RAM. In all experiments, we set K to the number of classes in the data set for all techniques except ICACLUS. ICACLUS does not require the number as clusters as input parameter but two other parameters: k specifying a number of relevant ICs and t specifying a similarity threshold. In all experiments, both were set to 1. We tried a range of different choices which all lead to worse results with excessive numbers of clusters. In addition, we set the maximum number of iterations for IKM to 75 in all experiments. For the efficient K-means-style methods IKM and SF, we used 100 random initializations and report the quality of the best result according to the internal objective function of the algorithm. Experiments with SCRA could only be performed with a single initialization due to excessive runtime, since the algorithm need hours to days to process a single data set. Experiments have been performed on 6 synthetic and 4 real-world data sets from various domains, for a summary see Table 1.

7.2 Effectiveness

To evaluate effectiveness, we report three established measures for clustering quality: the Rand Index (RI) [44] which is based on counting pairs of objects in the same class and in the same cluster and pairs of objects in different classes and clusters, respectively. Cluster Purity (CP) as introduced in [16] is the ratio of objects of the majority class of a cluster with respect to the size of the majority class. This is averaged among all clusters. Since RI and CP favor clusterings with imbalanced cluster sizes, we additionally report the Information Criterion (IC) [45]. IC is defined as the empirical conditional entropy between class- and cluster labels. Intuitively, IC corresponds to the number of bit required to encode the class labels of all objects given the cluster labels. For RI and CP, a good clustering obtains a high score. For IC a small value indicates a good clustering.

7.2.1 Synthetic Data

Proof of concept. We generated a synthetic data set ($DS1$) with 600 objects and 13 dimensions where each dimension is a timeseries with 3,333 time points. This data set contains six clusters. Each cluster consists of 100 objects which share a common interaction pattern as illustrated in Fig. 1. In this data set, for each cluster four dimensions are interacting. The interacting dimensions are perfectly correlated with $R^2 = 1$. The times series of 12 dimensions (among them the four interacting ones) have been generated according to the random walk model. The remaining dimension is uniformly distributed. As expected, IKM perfectly clusters this data set, with optimal scores of all quality measures see Table 2 and outperforms all comparison methods by a large margin. The second best result in terms

TABLE 2
Results on Benchmark Data

data set	method	RI	IC	CP
$DS1$ Synthetic	IKM	**0.99**	**0.09**	**99%**
	SF	0.49	1.48	18%
	ICACLUS	0.77	2.54	10%
	SCRA	0.17	2.58	14%
	Naive	0.72	2.38	30%
$DS2$ Synthetic w/o Interaction	IKM	0.77	2.54	10%
	SF	**1**	**0**	**100%**
	ICACLUS	0.97	0.15	64%
	SCRA	0.61	1.66	29%
	Naive	**1**	**0**	**100%**
$DS5$ EEG	IKM	**1**	**0**	**100%**
	SF	0.61	0.69	75%
	ICACLUS	0.51	0.3	10%
	SCRA	0.47	1.0	50%
	Naive	0.49	0.95	60%
$DS6$ CAD	IKM	0.91	0.91	67%
	SF	0.92	0.95	73%
	ICACLUS	0.85	2.21	19%
	SCRA	0.88	1.21	61%
	Naive	**0.98**	**0.2**	**90%**
$DS7$ Japanese vowels	IKM	**0.88**	**1.21**	**61%**
	SF	0.79	2.36	33%
	ICACLUS	0.79	2.88	15%
	SCRA	0.11	3.13	10%
	Naive	0.83	2.04	40%

1. http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

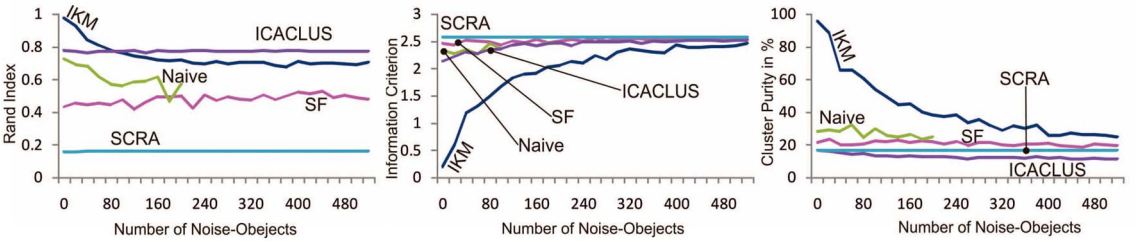

Fig. 4. Impact of noise objects on cluster quality: Noise with random labels.

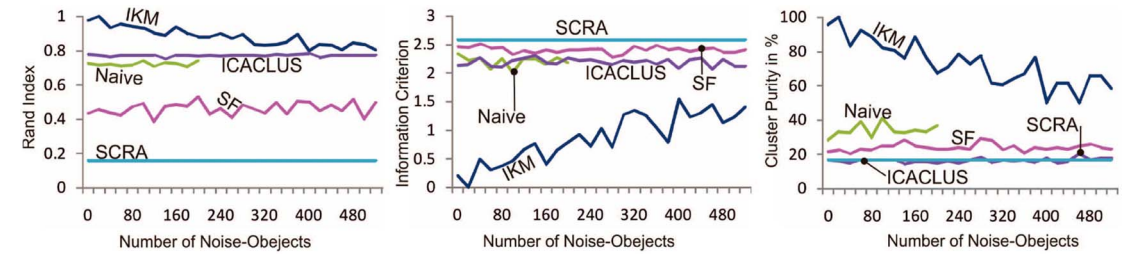

Fig. 5. Impact of noise objects on cluster quality: Accuracy without noise objects.

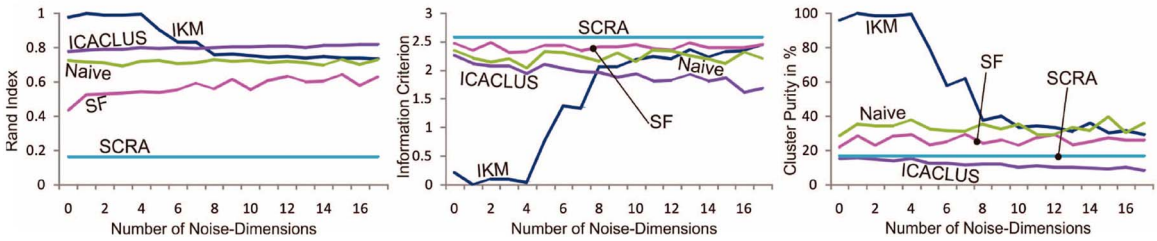

Fig. 6. Impact of noise dimensions on cluster quality.

of IC is obtained by SF (1.48), however this result has a RI of only 0.41 and CP of only 18%. The naive approach shows the worst IC of 2.38 but is somewhat better than SF in RI and CP. SCRA performs worst in all quality measures.

To demonstrate that IKM is designed to exploit interaction patterns for clustering, we created data set *DS*2 analogously to *DS*1 but without interactions among the dimensions. Instead, following the assumption of blind signal source separation, we created the data of each cluster as a unique linear mixture of independent sources including sine and sawtooth functions. On this data set, IKM is outperformed by all comparison methods in terms of CP and IC.

Robustness against noise objects. Real data often contain a large amount of noise objects which often severely affect the result of clustering techniques. In the cluster model of IKM, a noise object is an object without any interaction pattern. To systematically evaluate the performance of IKM in the presence of noise objects, we created a synthetic data set (*DS*3) similar to *DS*1 (six clusters, 12 dimensions, length of time series: 2,048) with various amounts of noise. Starting with 120 objects in six clusters without noise we successively added in each step 20 noise objects up to 520 noise objects. For evaluation, we assigned random class labels to the noise objects and report the clustering quality measures for the whole data set (cluster objects and noise objects, see Fig. 4)) and for the cluster objects only, see Fig. 5. As expected, all clustering quality measures worsen with increasing amount of noise objects for IKM. For SF, Naive, ICACLUS and SCRA, the clustering quality stays constant at a very poor level. It is natural that the clustering quality for the overall data displayed in Fig. 4 decreases when adding randomly labeled noisy objects. However, as Fig. 5 demonstrates, IKM achieves a correct clustering of the cluster points even in the presence of large amounts of noise and always performs superior than SF and Naive. With 80% of noise, IKM obtains more than 50% of CP (Fig. 5) or $RI \geq 0.8$ (Fig. 5). At 82% of noise IKM is still better then the SF-method and Naive without noise.

Robustness against noise dimensions. To study the robustness against noise dimensions, we successively added an increasing number of noise dimensions to our basic 12-dimensional data set (*DS*4). Noise dimensions have random values at all time points. Fig. 6 demonstrates that up to four noise dimensions do not affect the clustering accuracy of IKM, which corresponds to 25% of noise. More than 10 noise dimensions affect the result of IKM, but this means a noise percentage of 40% and more. The accuracy of the comparison methods remains constant, but at a pretty poor level. IKM outperforms them by a large margin up to 40% of noise and still remain better up to 60% of noise dimensions.

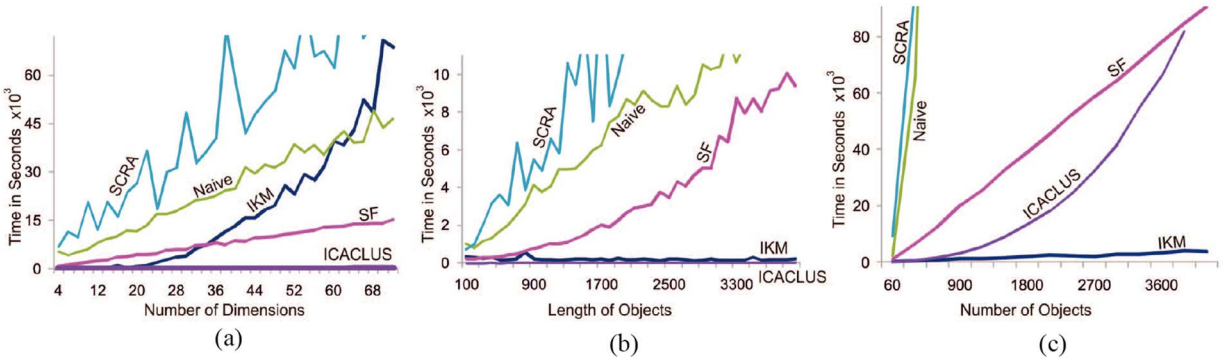

Fig. 7. Scalability w.r.t. dimensionality, time series length, and number of objects. (a) #dimensions/time. (b) length/time. (c) #objects/time

7.2.2 Real Benchmark Data

EEG data. Our first real-world data set *DS*5 is the EEG data set available at the UCI machine learning repository.[2] This data set is derived from a study on alternations of brain activity in alcoholic subjects in comparison to a control group. Each subject was exposed to either a single visual stimulus (S1) or to two visual stimuli (S1 and S2). When two stimuli were shown, they were presented in either a matched condition where S1 was identical to S2 or in a non-matched condition where S1 different from S2. Data set *DS*5 is derived from the so-called Small Data Set which contains data of 2 subjects, one alcoholic and one of class control. For each of the 3 experimental paradigms, 10 runs of EEG have been recorded. The task here is to cluster the EEG runs into alcoholic and control. IKM yields a perfect clustering with RI = 1, IC = 0 and CP of 100%. SF is the best performing comparison method with RI = 0.61, IC = 0.69 and CP of 75%. Besides the clustering, the result of IKM specifies the most important differences among the interaction patterns of the electrodes of the alcoholic subject and the healthy control. The most important electrodes for clustering are CZ (ch. 15), F8 (ch. 3), FT7 (ch. 36) and C3 (ch.16). The lateral and medial frontal brain areas under these electrodes can be associated with decision making processes. Substance-dependent individuals show disadvantageous decision-making, as well as altered frontocortical recruitment when performing experimental tasks [46], [47].

Common Activities data set. The data set *DS*6 from [40] consists of motion stream data of ten different activities performed by one person, for example pick up an object, wave or talk on a cell phone. There are ten different instances collected for each activity, hence 100 sequences in total. It is a non-trivial task to recognize the activities since they are typically performed with different rates of motion execution. Each activity is been represented by a multivariate times eries with 25 dimensions and 70 time points after Kernel PCA preprocessing with an RBF kernel [48]. On this data set, SF performs slightly better in Rand Index and Cluster Purity than IKM (see Table 2). The results reported for SF in [16] are even superior (only CP between 84% and 89%). We could not reproduce these results. Probably, the authors applied a different RBF-kernel for kernel PCA, which is unfortunately not exactly specified in [16]. However, the overall best result is achieved by the naive algorithm which is with 90% cluster purity even better than the results of SF reported in [16]. This demonstrates that our kernel PCA successfully captures the cluster separating information. In summary, we conclude, that interaction-based clustering is also successful in the application of motion stream data, but does not outperform SF and Naive.

Japanese Vowels. The real-world data set *DS*7 consists of two Japanese vowels uttered by nine different male speakers and is available at the UCI machine learning repository[3]. The task is to recognize the speakers based on 12-dimensional time series consisting of the Cepstrum coefficients of the speech signals. Representing speech by Cepstrum coefficients is wide-spread in language processing and speech recognition [41]. Cepstrum coefficients model the time evolving signal as an ordered set of coefficients representing the signal spectral envelope. The special characteristics of this data set are that time series vary very much in length (between 7 and 29 time points). We merge training set and test set for clustering. The results (see Table 2) demonstrate that IKM is the best choice to cluster the speakers since it clearly outperforms all comparison methods.

7.3 Efficiency

Table 1 provides a summary on the data sets $DS8 - DS10$ used for evaluation of scalability. These synthetic data sets are generated similar to *DS*1 varying all important properties. *DS*8 has 120 objects with a dimensionality ranging from 4 to 80. *DS*9 has 120 objects of the length from 100 to 4,000 with 100 as increment. *DS*10 consists of different numbers of objects. Let us note that the runtime of techniques coded in different languages is of course not directly comparable. However, all comparison methods are implemented in high-level programming languages and whenever available we used the original implementations provided by the authors. This runtime comparison intends to give a general impression about the scalability of the methods under comparison. We run all methods three times and report the average runtime.

Fig. 7(a) shows scalability in the dimensionality of the time series. For this example, the increase in runtime is

2. http://archive.ics.uci.edu/ml/databases/eeg/eeg.data.html

3. http://archive.ics.uci.edu/ml/datasets/Japanese+Vowels

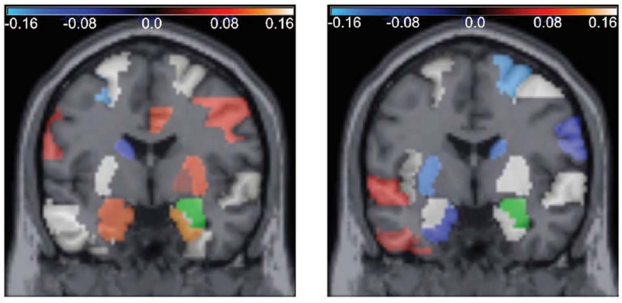

Fig. 8. Graphic representation of the models for the right amygdala (green) in patients and controls involving all 90 AAL-ROIs. While healthy controls show functional connectivity of the amygdala to sensory areas (superior temporal, auditory), patients show increased connectivity to frontal control areas (superior frontal). Red to white: Areas with a signal-time course with positive linear coefficient in the model, Blue: Negative linear coefficient, respectively.

TABLE 3
Results on fMRI Data

data set	method	RI	IC	CP
$DS11$ Somatoform	IKM	**0.56**	**0.89**	**69%**
	SF	0.48	1	50%
	ICACLUS	0.53	0.18	9%
	SCRA	0.48	1	50%
	Naive	0.49	0.98	58%
$DS11$ Somatoform after user interaction	IKM	**0.92**	**0.20**	**96%**
	SF	0.48	0.99	54%
	ICACLUS	0.54	0.18	35%
	SCRA	0.48	1.0	50%
	Naive	0.48	1.0	50%
$DS12$ Schizophrenia	IKM-linear	0.52	0.92	65.38%
	IKM-nonlinear	**0.73**	**0.62**	**84.62%**
	SF	0.49	0.92	57.69%
	Naiv	0.48	1.00	53.85%
	ICACLUS	0.52	0.83	27.69%
	SCRA	0.48	1	50%

approximately $O(d^3 \cdot \log(d))$ for IKM. ICACLUS, SF and Naive scale linearly in the dimensionality, where ICACLUS is most efficient on all data sets. SCRA is the least efficient method in comparison.

Fig. 7(b) shows scalability the length of the time series. Although the complexity is in principle linear for IKM, it can be very well amortized (we do not see time-growing). Similar to IKM, ICACLUS also scales very well. SF needs $O(d^2)$ of CPU-time. Naive is linear, but slower on our test data. Again, SCRA is least efficient. The high variation in runtime of SCRA is due to the fact that the algorithm depends very much on the initalization.

The Fig. 7(c) shows scalability in number of objects. The complexity is linear for IKM, SF, Naive and SCRA. ICACLUS scales quadratic in the number of objects. Nevertheless, the naive algorithm and SCRA are the slowest methods on all test data sets.

8 Interaction Among Brain Regions

8.1 Functional Magnetic Resonance Imaging

We obtained data sets $DS11$ and $DS12$ from functional MRI experiments. Functional MRI generates a series of 3-D volume images of the brain. Each image consists of about 60,000 voxels and the interval between time points is about 2-3 seconds. We first applied standard pre-processing including realignment, normalization to a standard template and smoothing. Our approach is based of a set of time-series. Basically we can use each voxel time series from the images. However, for neighboring voxels signal activity is very similar. Moreover, medical experts often desire to obtain results at the level of anatomical regions which facilitates interpretation. Therefore, in Section 8.2, we use a brain atlas from [49] with a predefined mask of regions. As an alternative to anatomical regions, in Section 8.3 we use Independent Component Analysis (ICA). From the ICA result, we selected physiologically relevant components and rejected ICs reflecting motions-artifacts or noise.

8.2 Somatoform Pain Disorder

$DS10$ [42] has been obtained from a study on Somatoform Pain Disorder and consists of images of 13 subjects with pain disorder and 13 healthy controls. Somatoform Pain Disorder has severe impact on the quality of living of the affected persons since the main symptom is severe and prolonged pain for which there is no medical explanation. The causes of this psychiatric disorder are not fully understood but the hypothesis is that patients have altered mechanisms of observing and processing pain. Therefore, in our experiment, subjects underwent alternating blocks of pain- and non-painful stimulation while in the scanner. After pre-processing we segmented the data of each subject into 90 anatomical regions of interest [49] (ROIs). The task is to cluster persons based on the interaction patterns of the ROIs within the brain during the experiment. Each person is represented by a multivariate time series with 90 dimensions and 325 time points. There are four subjects with 216 time points only. Our technique IKM does not require the multivariate time series subjected to a cluster analysis to be of equal length. For Naive and ICACLUS we can only use 216 time points for clustering, which implies a considerable information loss. A cluster analysis with IKM provides valuable insights into disordered brain connectivity of patients with pain disorder. Table 3 shows the clustering results. The result of IKM is superior to the results of all comparison methods: One cluster is composed of nine subjects with somatoform pain disorder and four healthy controls. The second cluster contains nine healthy controls and four subjects with somatoform pain disorder. Based on previous studies [42], [50], it is known that the right amygdala is strongly associated with somatoform pain disorder. The model of this region is the best separating model among the clusters. Fig. 8 presents a visualization of the model of right amygdala. The coefficients of this model are represented by color coding.

User Interaction. Based on the model displayed Fig. 8, our medical experts refined the set of ROIs in our input data to four regions in the orbitofrontal cortex: Inferior orbitofrontal (right and left) cortex, medial orbitofrontal cortex (right and left). These regions are also known to be involved in the representation of subjective feelings including pain [51]. After user interaction, IKM obtains a nearly perfect clustering: (see Table 3): Only one patient is

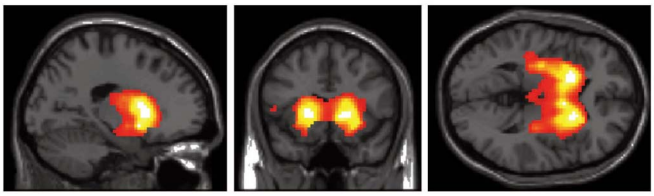

Fig. 9. Spatial map of the intrinsic basal ganglia network including the striatum.

put in a wrong cluster. The comparison methods do not profit much from the dimensionality reduction performed by our experts. This demonstrates that the cluster model of IKM successfully represents interaction patterns among brain regions. Furthermore, IKM yields interpretable results which can be improved by user interaction.

8.3 Schizophrenia

The dataset *DS*12 [43] is derived from 26 persons including 13 healthy controls and 13 patients with schizophrenia, who all were assessed by 10-minutes of resting-state functional MRI. Schizophrenia is characterized by the impaired interaction between distributed brain regions particularly the striatum. Increased dopamin activity in the striatum is essential for schizophrenia and antidopaminergic treatment the main therapy of the disorder. Therefore we suggested that the causal influence among intrinsic brain networks including the striatum is aberrant in patients. Intrinsic brain networks are characterized by synchronous brain activity at rest. (see Fig. 9 for an example: spatial map of the intrinsic basal ganglia network including the striatum). Independent component analysis of fMRI data resulted in 9 ICs representing intrinsic brain networks by spatial maps of synchronous activity and corresponding time series. These time series result in 26 multidimensional time series objects of healthy controls and patients. Causal influence from one area into another was modeled by Granger Causality between time series of brain network activity.

Clustering based on nonlinear models reflecting Granger causality separated patients from controls with high cluster purity (84.6%) consistently for a model of the striatum. Each cluster consists of 13 persons. In total, only two persons (one control and one patient) have been incorrectly clustered. Changed influence on the striatum was found for several intrinsic brain networks, indicating an aberrant regulation of striatal activity. These data suggest that altered regulatory intrinsic network activity contributes to increased striatal dopamin function.

Table 3 shows the results of nonlinear and linear model-based IKM and the results of the comparison methods for *DS*12. Nonlinear IKM clearly outperforms all comparison methods.

9 Conclusion

In this paper, we propose a novel cluster notion for multivariate time series. We define a cluster as a set of objects sharing a specific interaction pattern among the dimensions. In addition, we propose interaction K-means (IKM), an efficient algorithm for interaction-based clustering. Our experimental evaluation demonstrates that the interaction-based cluster notion is a valuable complement to existing methods for clustering multivariate time series. IKM achieves good results on synthetic data and on real world data from various domains, but especially excellent results on EEG and fMRI data. Our algorithm is scalable and robust against noise. Moreover, the interaction patterns detected by IKM are easy to interpret and can be visualized. Nonlinear models show their superiority in the corresponding real world data. In ongoing and future work, we plan to extend our ideas to differential equations. We want to consider different models for different regions of the time series. We intend to work on methods for suitable initialization of IKM, since existing strategies for K-means can not be straightforwardly transferred to IKM because of the special cluster notion. We are also investigating in feature selection for interaction-based clustering.

Acknowledgments

The authors thank M. Valet, C. Zimmer, T. Sprenger, A. Backrests, T. R. Tölle, and H. Gündel for providing us the fMRI data. They additionally thank X. Wang, A. Wirth, and L. Wang (the authors of [16]) for providing us the source code of their method. C. Plant is supported by the Alexander von Humboldt-Foundation.

References

[1] M. D. Fox and M. E. Raichle, "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging," *Nat. Rev. Neurosci.*, vol. 8, no. 9, pp. 700–711, 2007.

[2] C. Sorg *et al.*, "Selective changes of resting-state networks in individuals at risk for alzheimer's disease," *PNAS*, vol. 104, no. 47, pp. 18760–18765, 2007.

[3] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Boston, MA, USA: Elsevier, 2007.

[4] S. Smith *et al.*, "Correspondence of the brain's functional architecture during activation and rest," *PNAS*, vol. 106, no. 31, p. 13040, 2009.

[5] C. Li, L. Khan, and B. Prabhakaran, "Feature selection for classification of variable length multiattribute motions," in *Multimedia Data Mining and Knowledge Discovery*, V. A. Petrushin and L. Khan, Eds. London, U.K.: Springer, 2007.

[6] C. Pamminger, *Bayesian Clustering of Categorical Time Series: An Approach Using Finite Mixtures of Markov Chain Models*. Saarbrücken, Germany: VDM Publishing Group, 2008.

[7] T. W. Liao, "Clustering of time series data – A survey," *Pattern Recognit.*, vol. 38, no. 11, pp. 1857–1874, 2005.

[8] E. Keogh and S. Kasetty, "On the need for time series data mining benchmarks: A survey and empirical demonstration," in *Proc. SIGKDD*, Edmonton, AB, Canada, 2002, pp. 102–111.

[9] J. Lin, M. Vlachos, E. Keogh, and D. Gunopulos, "Iterative incremental clustering of time series," in *Proc. EDBT*, Heraklion, Greece, 2004, pp. 106–122.

[10] M. Vlachos, J. Lin, E. Keogh, and D. Gunopulos, "A wavelet-based anytime algorithm for k-means clustering of time series," in *Proc. Workshop Clustering High Dimensionality Data Its Applications*, 2003, pp. 23–30.

[11] T. Oates, L. Firoiu, and P. R. Cohen, "Clustering time series with hidden Markov models and dynamic time warping," in *Proc. IJCAI Workshop Neural, Symbolic Reinforcement Learning Methods for Sequence Learning*, 1999, pp. 17–21.

[12] T. Oates, "Identifying distinctive subsequences in multivariate time series by clustering," in *Proc. ACM SIGKDD*, New York, NY, USA, 1999, pp. 322–326.

[13] J. Lin *et al.*, "A MPAA-based iterative clustering algorithm augmented by nearest neighbors search for time-series data streams," in *Proc. PAKDD*, Hanoi, Vietnam, 2005, pp. 333–342.

[14] C. A. Ratanamahatana, E. J. Keogh, A. J. Bagnall, and S. Lonardi, "A novel bit level time series representation with implication of similarity search and clustering," in *Proc. PAKDD*, Hanoi, Vietnam, 2005, pp. 771–777.

[15] X. Ge and P. Smyth, "Deformable Markov model templates for time-series pattern matching," in *Proc. KDD*, New York, NY, USA, 2000, pp. 81–90.

[16] X. Wang, A. Wirth, and L. Wang, "Structure-based statistical features and multivariate time series clustering," in *Proc. ICDM*, Omaha, NE, USA, 2007, pp. 351–360.

[17] J. Lin and E. Keogh, "Clustering of streaming time series is meaningless," in *Proc. 8th ACM SIGMOD Workshop Research Issues*. New York, NY, USA: ACM, 2003, pp. 56–65.

[18] E. H. C. Wu and P. L. H. Yu, "Independent component analysis for clustering multivariate time series data," in *Proc. ADMA*, Wuhan, China, 2005, pp. 474–482.

[19] L. Owsley, L. Atlas, and G. Bernard, "Automatic clustering of vector time-series for manufacturing machine monitoring," in *Proc. IEEE ICASSP*, vol. 4. Munich, Germany, 1997, pp. 3393–3396.

[20] X. Z. Wang and C. McGreavy, "Automatic classification for mining process operational data," *Ind. Eng. Chem. Res.*, vol. 37, no. 6, pp. 2215–2222, 1998 [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ie970620h

[21] C. Plant, A. M. Wohlschläger, and A. Zherdin, "Interaction-based clustering of multivariate time series," in *Proc. ICDM*, Miami, FL, USA, 2009, pp. 914–919.

[22] F. Mörchen, "Time series feature extraction for data mining using DWT and DFT," Dept. Math. Comput. Sci., University of Marburg, Marburg, Germany, Tech. Rep. 33, 2003.

[23] A. Sudjianto and G. S. Wasserman, "A nonlinear extension of principal component analysis for clustering and spatial differentiation," *IIE*, vol. 28, no. 12, pp. 1023–1026, 1996.

[24] A. Trouve and Y. Yu, "Unsupervised clustering trees by nonlinear principal component analysis," *Pattern Recognit. Image Anal.*, vol. 2, pp. 108–112, 2001.

[25] E. J. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in *Proc. KDD*, New York, NY, USA, 2004, pp. 206–215.

[26] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[27] H.-P. Kriegel, P. Kröger, A. Pryakhin, M. Renz, and A. Zherdin, "Approximate clustering of time series using compact model-based descriptions," in *Proc. DASFAA*, New Delhi, India, 2008, pp. 364–379.

[28] C. Böhm, L. Läer, C. Plant, and A. Zherdin, "Model-based classification of data with time series-valued attributes," in *Proc. BTW*, 2009, pp. 287–296.

[29] P. Cheeseman and J. Stutz, "Bayesian classification (autoclass): Theory and results," in *Proc. Adv. Knowl. Discov. Data Min.*, Menlo Park, CA, USA, 1996, pp. 153–180.

[30] A. R. McIntosh and B. Miic, "Multivariate statistical analyses for neuroimaging data," *Annu. Rev. Psychol.*, vol. 64, no. 1, pp. 499–525, 2013 [Online]. Available: http://www.annualreviews.org/doi/abs/10.1146/annurev-psych-113011-143804

[31] D. T. Larose, *Data Mining Methods and Models*. Hoboken, NJ, USA: Wiley, 2006.

[32] E. I. George, "The variable selection problem," *J. Amer. Statist. Assoc.*, vol. 95, no. 452, pp. 1304–1308, 2000.

[33] E. Formisano and R. Goebel, "Tracking cognitive processes with functional mri mental chronometry." *Current Opinion Neurobiol.*, vol. 13, no. 2, pp. 174–181, 2003 [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0959438803000448

[34] J. Hamilton, G. Chen, M. Thomason, M. Schwartz, and I. Gotlib, "Investigating neural primacy in major depressive disorder: Multivariate granger causality analysis of resting-state fmri time-series data," *Mol. Psychiatry*, vol. 16, no. 7, pp. 763–72, 2011.

[35] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symposium Mathematical Statistics and Probability*, vol. 1, L. M. L. Cam and J. Neyman, Eds. Berkeley, CA, USA: University of California Press, 1967, pp. 281–297.

[36] D. Arthur, B. Manthey, and H. Röglin, "Smoothed analysis of the k-means method," *J. ACM*, vol. 58, no. 5, pp. 19:1–19:31, Oct. 2011 [Online]. Available: http://doi.acm.org/10.1145/2027216.2027217

[37] Y. Chen *et al.*, "Regression cubes with lossless compression and aggregation," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 12, pp. 1585–1599, Dec. 2006.

[38] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, 2000.

[39] L. Xiao, H. Begleiter, B. Porjesz, W. Wenyu, and A. Litke, "Event related potentials during object recognition tasks," *Brain Res. Bull.*, vol. 38, pp. 531–538, 1995 [Online]. Available: http://www.ingentaconnect.com/content/els/03619230/1995/00000038/00000006/art02023

[40] A. Veeraraghavan and A. K. R. Chowdhury, "The function space of an activity," in *Proc. CVPR*, New York, NY, USA, 2006, pp. 959–968.

[41] M. Kudo, J. Toyama, and M. Shimbo, "Multidimensional curve classification using passing-through regions," *Pattern Recognit. Lett.*, vol. 20, no. 11-13, pp. 1103–1111, 1999.

[42] H. Gündel *et al.*, "Altered cerebral response to noxious heat stimulation in patients with somatoform pain disorder," *Pain.*, vol. 137, no. 2, pp. 413–421, Nov. 2007.

[43] C. Sorg *et al.*, "Increased intrinsic brain activity in the striatum reflects symptom dimensions in schizophrenia," *Schizophr Bull.*, vol. 39, no. 2, pp. 387–395, 2012.

[44] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *J. Intell. Inf. Syst.*, vol. 17, no. 2–3, pp. 107–145, 2001.

[45] B. Dom, "An information-theoretic external cluster-validity measure," IBM Research Division, Tech. Rep. RJ 10219, May 2001 [Online]. Available: citeseer.nj.nec.com/dom01informationtheoretic.html

[46] H. Heekeren, S. Marrett, and L. Ungerleider, "The neural systems that mediate human perceptual decision making," *Nat. Rev. Neurosci.*, vol. 9, no. 6, pp. 467–79, Jun. 2008.

[47] G. Fein and M. Chang, "Smaller feedback ERN amplitudes during the BART are associated with a greater family history density of alcohol problems in treatment-naïve alcoholics," *Drug Alcohol Depen.*, vol. 92, no. 1–3, pp. 141–8, Jan. 2008.

[48] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[49] N. Tzourio-Mazoyer *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, Jan. 2002.

[50] I. Strigo, A. Simmons, S. Matthews, A. Craig, and M. Paulus, "Association of major depressive disorder with altered functional brain response during anticipation and processing of heat pain." *Arch. Gen. Psychiatry*, vol. 65, no. 11, pp. 1275–84, Nov. 2008.

[51] M. L. Kringelbach, "The human orbitofrontal cortex: Linking reward to hedonic experience," *Nat. Rev. Neurosci.*, vol. 6, no. 9, pp. 691–702, 2005.

Claudia Plant received her PhD degree in 2007 and currently is leading the research group iKDD (integrative Knowledge Discovery in Databases) at Helmholtz Zentrum München and Technische Universität München, Munich, Germany. Her research focuses on database-related data mining, especially clustering, parameter-free data mining and integrative mining of heterogeneous data. She not only contributed to top-level database and data mining conferences like KDD, SIGMOD, ICDM, ICDE but also published application-related data mining approaches resulting from interdisciplinary projects with experts from biology, neuroscience and environmental engineering in leading journals like *Bioinformatics*, *Cerebral Cortex*, and *Water Research*.

Andrew Zherdin received his degree in informatics in 2007. Currently, he works as a scientist in the Department of Neuroradiolgy at Technische Universität München TUM and he wrote his PhD-thesis. His research focuses on database-related data mining, clustering, multivariate times eries and mathematical models. He has published several papers in the field of data-mining and neuroscience.

Christian Sorg received his master's degree in mathematics in 1998 and his MD degree in 2008. Currently, he works as a psychiatrist in the Department of Psychiatry and Neuroradiolgy at Technische Universität München TUM and he is leader of the research group Large-scale Brain Models of Psychiatric Disorders at the Neuroimaging Center NIC TUM. His research focuses on the multimodal brain-connectivity-based models of distinct psychiatric disorders particularly Alzheimer's disease, schizophrenia, and borderline personality disorder. He has published in several high-rank journals in the field of clinical and general neuroscience such as *Biological Psychiatry*, *Current Alzheimer Research*, and *PNAS.*

Anke Meyer-Baese is an Associate Professor in the Department of Scientific Computing at Florida State University. She received her PhD degree in Electrical Engineering from Darmstadt University of Technology in 1995. Her research interests lie in the area of signal processing and data mining with emphasis on medical applications. Her publication record consists of over 140 contributions to leading journals and conferences, several book chapters and two books.

Afra M. Wohlschläger received her PhD degree in Physics from the University of Cologne in 1997. Currently, she is the leader of the research group Functional Connectivity of the Brain at Rest at the Department of Neuroradiology at Technische Universität München (TUM). Her former affiliations include the Research Center Jülich, Germany, and the Ben-Gurion University of the Negev, Israel. Her research focuses on long-term changes Intrinsic brain networks (BNETs) in neuropsychiatric diseases (resting-state fMRI), experience-dependent plasticity of BNETs in healthy subjects and patients, structural brain networks (DTI) and data mining algorithms for neuroimaging data. She has published in leading journals including the *Journal of Neuroscience, Neuroimage*, and *Science.*